The Use of Syntactic Structure in Relationship Extraction

by

Natasha Singh

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2004

Author .....
Department of Electrical Engineering and Computer Science
May 20, 2004

Certified by .....
Michael J. Collins
Assistant Professor
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Theses

The Use of Syntactic Structure in Relationship Extraction

by

Natasha Singh

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2004, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

This thesis describes a method of relationship extraction that uses surface and syntactic features of text along with entity detection information to perform sentence-level relationship extraction tasks. The tasks are to build and test classifiers for the following relationships: employer–employee, organization–location, family, and person–location. Methods of reducing noise in these features sets are also discussed, and experimental evidence of their effectiveness is presented. Performance of the system is analyzed in terms of precision and recall, and errors of the system are discussed along with proposed solutions. Finally a reformulation of the problem is presented along with a discussion of future work.

Thesis Supervisor: Michael J. Collins
Title: Assistant Professor

Acknowledgments

My advisor Prof. Michael Collins has been an invaluable source of guidance throughout this project. It is has truly been a privilege working with him, and I look forward to continuing my education with him.

Furthermore I would like to thank each of my group members, especially Percy for his helpful insights, and Luke for his candid reviews. I would also like to thank the rest of my group for the ideas they inspired, and their support.

I would like to thank my parents for their continuing emotional and financial support over the many years I have been a student. They provided me with the freedom and attitude to pursue my education, and for that I am very grateful.

Finally I would like to thank my friends and sisters for their support and making life fun. I would especially like to thank Nick for his continued support and providing a fresh unbiased mind for my ideas.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Communicating information through language is critical to our ability to interact and function successfully in the world. We utilize language to communicate many types of information. Descriptions of entities and the way in which they are related are one form of information that is conveyed. As we strive to create more intelligent computer programs that are more adept at interacting with people, it is important that we develop the ability of machines to be able to detect entity relationship information. The goal of this thesis is to investigate a method emphasizing the use of syntactic information to extract relations between entities from text.

This thesis intends to study how relationships are expressed in single sentences, rather than across multiple sentences or within a whole document. Sentences were chosen because they are a fundamental unit that can contain a relationship. Once a reliable way of detecting relationships at this level is deveoped, cross sentence and document models can be introduced to obtain a comprehensive relationship extraction system.

There are a number of compelling reasons for studying relationship extraction (besides the satisfaction of intellectual curiosity). Relationship extraction technology could allow for automated detection of relations between entities. In this manner large databases of relational information can be accumulated, generating new information for data mining, question answering systems, and information retrieval (search).

This thesis is structured in the following manner. First the field of information extraction will be described, followed by a description of previous work in relationship extraction. Following that a description of the data set and the method of extraction will be discussed. Finally experimental results and a thorough error analysis will be presented, along with a discussion of future work in this area.

1.1 Information Extraction (IE)

The goal of **information extraction** is to accumulate semantic information from text. This usually involves first spotting the important entities (people, organizations, locations, etc.), and then discovering the way in which they are related. For example, consider the following sentence.

```
Mr.  Smith, CEO of ACME Corporation, leaves for Washington tomorrow.
```

An information extraction system should recognize that "Mr. Smith" is a person, "ACME Corporation" is an organization, and "Washington" is a location. This task is known as **named entity detection**. Relationships or events can also be extracted from this sentence. For example, "Mr. Smith" is CEO of "ACME Corporation", "Mr. Smith" works for "ACME Corporation", and "Mr. Smith" will visit "Washington". These examples are depicted in Table 1.1. Populating such tables is the goal of **relationship extraction** and **event extraction**. Notice that in the table two relationships–"Mr Smith" works for "ACME Corporation" and "Mr. Smith" is CEO of "ACME Corporation"–are combined into a single entry.

1.2 Relationship Extraction

Relationships are one type of data that can be extracted from text. Minimally, a relationship is some association between two entities. A relationship might also involve multiple entities, or be characterized by other attributes. The tense of a relationship (past, present, future) is an example of such an attribute. Relationships

Table 1.1: Sample tabular representation of extracted relationships and events.

EVENT	PERSON will go to LOCATION
PERSON	Mr. Smith
LOCATION	Washington

RELATIONSHIP	PERSON works for ORGANIZATION as ROLE
PERSON	Mr. Smith
ORGANIZATION	ACME Corporation
ROLE	CEO

may also be related hierarchically, for example, a person who is "CEO of" Company A, also "works for" the Company A.

In this thesis, the following four simple binary relationships are considered, *employer–employee*, *family relationship*, *organization–location*, and *person–home*. Each relationship is defined and an example of each is provided in Table 1.2. A detailed specification of these relationships is provided in Chapter 3. These relationships were selected because they seem to be common as well as useful relationships to extract. However, it is important to note that the definition of each relationship may have be subjective. While attempting to build an extraction system, it may be discovered that different ways of defining the structure of the desired information may provide less subjective boundaries.

Table 1.2: Relationships considered in this thesis. Examples were drawn from New York Times News Service [11]

Relationship	Definition	Example
employer employee	*PERSON* works/worked/will work for *ORGANIZATION*	*Jack Dongarra*, a computer scientist at the *University of Tennessee* who maintains an annual list of the world's 500 fastest computers, finds that about 250 machines fall off his list yearly.
family relationship	*PERSON* has/had/will have a family relationship with *PERSON*	But *Joanne Goldwater*, the senator's daughter, and *Bob Goldwater*, his brother, said no plans had been made.
organization location	*ORGANIZATION* is/was/will be located in *LOCATION*	*Central Investment*, which also bottles Pepsi products in *Fort Lauderdale* and Palm Beach, Fla., demanded more territory in Florida and Ohio, money for marketing expenses and a limit on prices of concentrate, Pepsi officials said.
person location	*PERSON* lives/lived/will live in *LOCATION*	Two other candidates, *Lewis Leslie McAllister Jr.*, a *Tuscaloosa* businessman, and Phillip Williams, a former state finance director, shared the rest of the vote.

Chapter 2

Background

2.1 Message Understanding Conference (MUC)

The state-of-the-art IE systems were put to the test at the MUC competitions during the 1990's. This conference helped to set goals for the IE community and to guide the direction of research in this area. At the last conference, MUC-7, in 1998 [23], systems competed in the areas of information extraction depicted in Table 2.1. These systems were tested and trained on New York Times News Service data. Three specific relationships–" employee_of", "manufacture_of", and "location_of"– were considered in this competition. This work helps to set a foundation of the structure and type of relationships to focus on in this thesis. While the relationships considered in the competition were organization-centric, different varieties of entity pairs are considered in this thesis.

2.2 Evaluation Metrics

Relationship extraction is often measured in terms of how many of the relationships are successfully extracted (recall or coverage), and the proportion of extracted relationships that are accurate (precision). Another metric, the F-measure is a combina-

Table 2.1: MUC-7 IE tasks, table recreated from [14].

Template Relation Task	Extract relational information on employee_of, manufacture_of, and location_of relations
Scenario Template Task	Extract prespecified event information and relate the event information to particular organization, person, or artifact entities involved in the event
Coreference Task	Capture information on coreferring expressions: all mentions of a given entity, including those tagged in NE, TE tasks

tion of these measures. These are defined in Equation 2.1.

$$\begin{aligned} A &= \#\text{ of relationships in the text} \\ B &= \#\text{ of relationships extracted} \\ C &= \#\text{ of correct relationships extracted} \\ \text{Recall } (R) &= \frac{C}{A} \\ \text{Precision } (P) &= \frac{C}{B} \\ \text{F-measure } (F) &= \frac{RP}{\frac{1}{2}(R+P)} \end{aligned} \tag{2.1}$$

2.3 Successful Approaches to Relationship Extraction

There have been several successful approaches to relationship extraction over the last several years. These approaches have had varying goals. A few of them, DIPRE, Snowball, FASTUS, and SIFT, are discussed below. DIPRE aimed to extract at least one mention of each relationship efficiently from the World Wide Web, and thus complete coverage was not as important as efficiency. Snowball aimed to extract all mentions of a relationship with high coverage from large text collections, requiring a high degree of efficiency as well as language understanding. FASTUS also aimed to extract all relationship mentions from a large text corpus. SIFT attempted to extract relationships with high precision and recall from smaller text collections, allowing

the system to perform in-depth language analysis. The goals of these systems are summarized in Table 2.2.

Table 2.2: The goals of a few successful relationship extraction systems: DIPRE, Snowball, and SIFT.

	Large Corpora	**Smaller Corpora**
Coverage of Every Mention	Snowball, FASTUS	SIFT
Coverage of Every Relationship	DIPRE	

Systems have also used important techniques such as bootstrapping, where learning of relationship models is done iteratively (see description of DIPRE below). Also some have used integrated strategies where relationship extraction is done concurrently with other steps, such as entity tagging or sentence parsing. While all these techniques have boasted some amount of success, none has produced a high precision, high coverage, generalized relationship extractor for either large or small text corpuses.

2.3.1 Dual Iterative Pattern Expansion (DIPRE)

One of the first successful relationship extraction systems was DIPRE developed by Sergey Brin [4] which uses an unsupervised approach. This system was designed to extract relational information from unlabeled text on the World Wide Web, requiring very little training. DIPRE uses the following bootstrapping approach.

(1) Start with a few examples of a certain relationship (e.g.*(Bill, Hilary)* for family relationship)

(2) Locate many occurrences of these examples (e.g. ‘‘`Bill's wife Hilary will be at the White House.`’’

(3) Form generalized patterns of the relationship from these occurrences

(4) Use these patterns to extract more examples

(5) Repeat

The patterns that define a relationship consist of the following tuple: *(order, urlprefix, prefix, middle, suffix)*. *Order* is a boolean value indicating which entity in the relationship occurred first, *urlprefix* refers to web address of the information, and *prefix, middle*, and *suffix* are the strings that occur before the first entity, between the two entities, and after the second entity. Patterns have a *specificity* measure to ensure that they are not too general (see [4] for details). There also may be some regular expression restrictions on the entities in the relationship.

This system provides an elegant method of extracting relationships using unlabeled data. Defining a new relationship requires the minimal effort of finding a few examples of the relationship. The simplicity of the patterns allows for very fast computation. Also, given that the World Wide Web is so large, relationships are probably mentioned several times. This means that only a small percentage of the relationships need to be spotted in order to cover all the necessary pairs. DIPRE is able to extract a great deal of useful information with minimal user input and computation

However this solution also has limitations. The first is that no information about the type of entity is used in this process. This information can be useful in avoiding incorrectly identifying a relationship pair. For example, consider the following two sentences.

```
Sheila Kingsley works on the NASA Space Shuttle Program.
American Computer works on the Computer-In-The-Classroom Project.
```

Only the first sentence indicates an employer–employee relationship. The second does not. The only way to know this is to recognize that "Sheila Kingsley" is a person's name, while "American Computer" is an organization's name.

Another limitation of this approach is that the simplicity of the patterns will prevent the system from identifying a high percentage of the relationship occurrences. This was not the goal of DIPRE, but this may be useful when using a smaller amount of text where more depth of understanding is desired, or when a relationship will not be mentioned very frequently.

2.3.2 Snowball

Snowball is a relationship extraction system developed be Eugene Agichtein and Luis Gravano in 2000 [1]. This system uses a bootstrapping approach similar to that of DIPRE, however this system aims to have increased coverage. A significant difference between Snowball and DIPRE is that Snowball uses an entity detector in order to locate possible related entities. Snowball also uses a similarity measure to compare the *prefix, middle* and *suffix* strings, and hence does not require that they be an exact match. This means that small added words, such as an extra adjective or determiner will not prevent a perfectly good relationship from being extracted.

In the test results reported in [1], Snowball performs significantly better than DIPRE in terms of both precision and recall. However, there are still large portions of data Snowball is not able to process correctly. While the patterns are slightly more lax in Snowball than in DIPRE, they still do not allow for extra phrases or several words to be inserted. Also the laxness of the pattern may allow for incorrect matching. Consider the two following sentences.

```
ACME Corporation will be hiring Sheila Kingsley.
ACME Corporation will not be hiring Sheila Kingsley.
```

Here we have a simple example where a single word "not" indicates that a relationship does not hold. Without some additional levels of modelling, detecting such cases is impossible.

2.3.3 Finite State Automaton Text Understanding System (FASTUS)

There have been several systems that have attempted to apply natural language technologies to perform relationship extraction. They generally adhered to the following serial strategy [17].

(1) Perform part-of-speech tagging

(2) Perform name finding

(3) Perform syntactic analysis (usually shallow parsing)

(4) Perform semantic interpretation (usually pattern matching)

The FASTUS system developed by Jerry Hobbs and David Israel is an example of this approach [12]. This system is a fast, commercial information extraction system designed for use by the intelligence community. The system follows three steps (recreated from [12]).

(1) Recognizing Phrases: Sentences are segmented into noun groups, verb groups, and other phrases.

(2) Recognizing Patterns: The sequence of phrases produced in Step 1 is scanned for patterns of interest, and when they are found, corresponding "incident structures" are built.

(3) Merging Incidents: Incident structures from different parts of the text are merged if they provide information about the same incident.

The "incident structures" in FASTUS are structures that can model both relationships and events. FASTUS is able to process an average length news article in a few seconds.

While this approach uses some linguistic knowledge, such as part-of-speech tagging and shallow syntax parsing, it does not make use of a full syntax parse. This may lead to problems, as the full parse is often helpful in determining the exact meaning of text.

2.3.4 SIFT

The problem with a serial approach to information extraction is that later steps are not allowed to influence earlier steps. For instance, consider the following sentence.

```
Fannie Mae is executive vice president of ACME Corporation.
```

"Fannie Mae" could refer to a person or to company. We know that it refers to a person's name here because of the employer–employee relationship expressed in the sentence. For this reason, BBN's SIFT system adopted a unified approach where all these steps were performed synchronously [16]. Also SIFT performed full syntax parsing as well as complex semantic parsing. The system is trained on hand tagged data from which it derives a statistical relationship extraction model.

SIFT achieved a recall of 64% and a precision of 81% on the MUC-7 relationship extraction task. These tests require the extraction of all relationship mentions from a set of newspaper articles.

One of the drawbacks to this approach is that it may be difficult to obtain sufficient training data to accurately extract a relationship. Also, once created, the system may be difficult to modify. Furthermore, the system may not easily extend to include a wide variety of relationships.

2.3.5 Discrepancies in Evaluating Methodology and Data Sets

Besides varying goals, many of the systems for relationship extraction have also varied in the method of testing. DIPRE was tested by studying twenty extracted relationships for quality [4]. This type of testing only allows for measuring precision, but not recall. Also, such a small amount of test data does not allow for significant determination of the accuracy of the system, provide examples that will help detect flaws with the system, or allow comparison between systems. Snowball tested for coverage and precision by using a database of known valid pairs as well as sampling [1]. SIFT used the MUC-7 hand labeled test data [16]. The difficulty with these last approaches is that is requires a fair amount of hand tagged data, especially when considering test multiple relationships. Also, if the testing data for different systems comes from different sources, they may vary a great deal. These variations often prevent direct comparison between different relationship extraction techniques.

2.4 Framing the Problem

The goal of this thesis is to perform an in-depth study of how several binary relationships are expressed within single sentences. The use of several relationships will be useful in discovering commonalities as well as differences across different relationships. In each sentence, every pair of the target entity types (e.g. (PERSON, LOCATION)), will be considered a separate **candidate**. A candidate consists of two entities e_1 and e_2 along with a sentence s, to form the triple (e_1, e_2, s). This thesis focuses on classifying each candidate as either positive (the relationship holds) or negative (the relationship doesn't hold). The framing of the problem as a binary classification problem follows from work performed by Dmitry Zelenko, Chinatsu Aone and Anthony Richardell [27].

2.5 Lessons Learned

Many important applications of relationship extraction require high precision and recall on a variety of text sources. Also this problem closely mimics the abilities of human beings to perform relationship extraction. For these reasons, the goal of this thesis is to achieve high precision and recall. Therefore the measure for success will be a high F-measure, as well as a large area under the precision vs. recall curve.

In order to accurately determine the precision and recall of the system, testing will occur on large amounts of hand-tagged data. This will provide confidence in the system as well as the ability to compare different approaches.

Chapter 3

Relationship Specifications

As stated in Chapter 1, four binary relationships are considered in this thesis: *employer–employee*, *family*, *person–home*, and *organization–location*. The training and test sets for these relationships are derived from the AQUAINT Corpus by David Graff [11].

These relationships were selected because they represent relationships between several entity categories (organization–person, person–person, person–location, and organization–location). They also have different relative frequencies in our training data (e.g. employer–employee is far more frequent that the family relationship). These relationships can be expressed in several ways, which means they are not trivial pieces of information to extract using simple rules or textual patterns. Finally, each relationship is specific enough so that cooccurrence of two entities would not be enough information to determine the relationship (knowledge of language is necessary). A detailed specification and several positive candidates of each relationship are presented below, along with a discussion of the relationship. The positive candidates are chosen randomly.

3.1 Specification

The following specification is used to tag training and test data for this thesis. The definition of each relationship is stated in Chapter 1. First a general specification is presented that applies all four relationships. Specialized rules for each relationship are discussed in their respective sections.

1. The relationship could have held at any point in time past, present, or future.

2. Speculation on a relationship should not be tagged positively, (e.g. ‘‘`Mary and Ben might get married`’’.)

3. The relationship must be stated within the sentence in question, and should not be inferred from other information. For instance, consider the sentence ‘‘`Andrew Jones, Bill Gates' assistant, spoke with reporters about Microsoft's current market share.`’’ This sentence does not indicate that Andrew Jones works for Microsoft unless one knows that Bill Gates works for Microsoft, therefore this example should be tagged negatively.

4. On the other hand, examples known to be negative from outside information should be tagged negatively. For example given the following sentence ‘‘`Thomas and Ben Smith both play for TeamX`’’ it may be logical to assume the Thomas and Ben are related. However, someone reading this sentence may be a huge fan of TeamX and therefore know that Thomas is a last name, not a first name, and that Thomas and Smith are in fact not related in any way.

5. Fictional information is valid if there are no clues in the sentence indicating the information is fictional. For instance `Dana Scully works for the FBI` is valid, but `Gillian Anderson plays an FBI agent` is not valid.

6. Only relationships between the two entities indicated, and not repeated mentions of the same entity in a sentence should be considered (each case is considered separately). For example, the underlined pair of entities are a postive example in the first of the following sentences, but not in the second.

‘‘IBM will hire Ms. Scott’’, said IBM spokesman Martha Kelly.

‘‘IBM will hire Ms. Scott’’, said IBM spokesman Martha Kelly.

7. Mistakes of the named entity detector should be tagged negatively. For example consider the sentence ‘‘Fannie Mae plans to open a location in New York’’. If the entity tagger were to tag Fannie Mae as a person instead of an organization, and this should be presented as a candidate for the person–home relationship, it should be tagged negatively.

8. Often one cannot be totally certain of whether the candidate is positive or negative (the sentence is ambiguous). In these cases, if it is reasonable to assume that the author intended for the relationship to hold (the ambiguity is a technicality), then the example should be tagged positively. Otherwise it should be tagged negatively.

3.2 Employer-Employee Relationship

3.2.1 Specification

1. Being a member of an organization does not count, unless the person could be considered employed or working for the organization. For instance, being a member of a church does not count, but being a member of Congress does.

2. The employer–employee relationship includes part-time work, volunteer work, playing for a team, etc.

3. Being a student of a school is not included in this relationship.

3.2.2 Examples

The following are ten randomly chosen positive candidates of this relationship.

(1) “The government isn’t allowed to search your body or belongings without good reason,” said *Arthur Spitzer*, legal director of the Washington chapter of the *American Civil Liberties Union*.

(2) The first lady, Hillary Rodham Clinton, and Secretary of *State Madeleine Albright* traveled to the outskirts of Beijing on Monday morning to visit a women's legal aid center that operates on the leading edge of China's legal system, tackling a wide range of issues affecting women _ from rape to job discrimination to family planning.

(3) Indiana Pacers guard Jalen Rose, *Orlando Magic* forward *Bo Outlaw* and Boston Celtic foward Antoine Walker are a few NBA stars expected to compete.

(4) "Ever since we heard about his arrest, we've been trying in every conceivable way to try to convince the Chinese they made a grave mistake," said Tim Rieser, an aide to *Patrick Leahy* of Vermont, the top-ranking Democrat on the *Senate Foreign Operations Committee*.

(5) *Gordon Granger*, the *Union Army*'s commander in Texas, landing in Galveston on June 19, 1865, and issuing a proclamation about the end of slavery _ as well as the end of the Civil War, which had actually ended two months earlier.

(6) It also set in motion a chain of events that led to Ozawa's dismissal of Richard Ortner as administrator of the TMC and the high-profile resignations of the *TMC*'s artistic director *Leon Fleisher* and faculty chairman Gilbert Kalish.

(7) Instead, there were repeated accolades for Baruch's president, *Matthew Goldstein*, who is leaving this month to take the helm at *Adelphi Universtiy* and who helped turn Baruch into one of the most-respected colleges at CUNY during his seven years as president.

(8) *Sherri Hand*, an agricultural statistician for the *Kansas Department of Agriculture*, said early projections for this year's crop were conservative.

(9) ..." asked *Myron Magnet*, editor of City Journal, a publication of the conservative *Manhattan Institute*, which urges the mayor to battle on against other urban scourges like earsplitting car alarms and rude public servants.

(10) Dr. *Julio Montaner*, who heads the AIDS program at the *University of British Columbia* in Vancouver, said that HIV therapy was so complex that even experts like him had difficulty keeping track of therapeutic advances.

3.2.3 Discussion

These examples illustrate how the employer–employee relationship is expressed. Often the role a person plays in an organization is the primary clue to the relationship's existence. These roles can vary widely. In the above examples we see the roles of "legal director", "Secretary", "forward", "commander", "artistic director", "agricultural statistician", and "editor". Clearly it would be impossible to learn every role that indicates this relationship. Part of the challenge in extracting this relationship will be to find a way to generalize the detection of roles seen in training data.

Also, in the above candidates we see several expressions that indicate the employer-employee relationship: "on the", "take the helm", and"heads". Again a challenge arises from the large variety of these expressions that are important indicators of this relationship.

There may also be arbitrary extra text not revelant to this relationship in the sentence. It is important to be able to filter out excess information. For example in the first sentence the phrase "of the Washington chapter" is an extra phrase whose presence does not effect the extraction of this particular relationship.

3.3 Familial Relationship

3.3.1 Specification

1. Family includes immediate family, in-laws, step-families, etc.

2. Romantic involvements such as boyfriend, girlfriend, gay partner, etc. are valid.

3. If it is reasonable to assume that two people belong in the same family because of a shared last name, they should be tagged as positive (e.g. ‘‘`Mary-Kate and`

`Ashley Olsen acted in ...''`).

3.3.2 Examples

The following are ten randomly chosen positive candidates of this relationship.

(1) "Our chef for the past 20 years, Boris Bassin, saw the need for a lot of these changes," said Ann Zabar, whose father, *Saul*, owns Zabar's with his brother *Stanley*.

(2) *Kathy Klink* considered breaking off her relationship with *Paul*.

(3) *Eldon*'s wife, *Netta* (TraceyAtkins), is a vision of period domesticity is a polka-dot dress and apron, like an ad from a 1940s copy of House and Garden come to life.

(4) He will be the first Belmost Stakes starter for the trainer *Jimmy Jerkens*, who was a long-time assistant to his father, *Allen Jerkens*, before going out on his own in September.

(5) "My father didn't change that much," said *Ty*'s mother, the eldest of *Barry Goldwater*'s children, when I spoke with her later.

(6) *Akinwande* secretly kept in touch with his mother, *Josephine*, by telephone calls and letters.

(7) President *John F. Kennedy* named his brother and campaign manager, *Robert Francis Kennedy*, attorney general in 1961.

(8) When *Habibie* took the office of vice president earlier this year, he appointed a younger brother, *Effendi (Fanny) Habibie*, to succeed him as chairman of the Batam Island Devlopment Authority, the island's main governing body.

(9) *Nancy*, the woman *Ken* married a year ago in June, sees photos of him then, eating his way to 478 pounds _ he's around 240 now _ and she almost cries.

(10) Not long after, *Hernandez*, the older brother of Florida Marlins pitcher *Livan Hernandez*, went to Costa Rica.

3.3.3 Discussion

Here, as was seen in the employer–employee relationship, there are several roles that are strong indicators of the existence of this relationship, such as "brother", "mother", "father", etc. There are also a number of other words indicative of the relationship, such as "married" and "relationship". However, there appears to be less variance in these words that with the employer–employee relationship. It is therefore likely that the training data will cover most of range of these roles.

Furthermore, there appears to be a great deal of pronoun usage in expressing this relationship. For example, the word "he" refers to "Habibie" in example 8. Performing this resolution is necessary to identify the correct family relationship. This characteristic is stronger here than in the employer–employee relationship, as will be seen in Table 3.2.

Finally, this relationship also appears to be mutual. It seems necessarily true that if A is related to B, B is also related to A.

3.4 Person-Home Relationship

3.4.1 Specification

1. Working in a location is a positive indicator of this relationship.

2. Being an elected official from a location is a positive indicator of this relationship.

3. Being born in a location is a positive indicator of this relationship.

4. Being married to someone in a location is a positive indicator of this relationship.

3.4.2 Examples

The following are ten randomly chosen positive condidates of this relationship.

(1) "It's been like training camp," said starting center *Greg Foster* (an *Oakland* native), "and everybody knows training camp isn't any fun."

(2) "It would take more than 30 vials just to make someone sick, and we don't keep more than one or two in the office," said Dr. *Edmond Griffin* of Dermatology Associates of *Atlanta*, who has been offering the injection for more than two years.

(3) *Cardinal John O'Connor*, the archbishop of *New York*, used one of his Sunday homilies last month to warn of the bill's possible harm to society.

(4) *Ohio*-born lyric coloratura soprano *Kathleen Battle*, 49, is known for an incredibly sweet voice and incredibly difficult temperament; ...

(5) In seeking to appeal to Hispanics, Republicans have turned to *San Mateo County* Supervisor *Ruben Barrales* as their candidate for state controller.

(6) His brother, *Jim*, a jockey agent in *California*, thinks the "Hee Haw" accent is partly a put-on, Mike's way of suckering business rivals.

(7) Last fall, *North Carolina* writer *Kaye Gibbons* went from being a critically acclaimed author to best selling celebrity.

(8) Carolyn Maloney of New York and *John Dingell* of *Michigan*, who said they were looking for a backup in case more substantial reform bills failed.

(9) "And Senate Majority Leader *Joseph Bruno*, Donohue's former boss and next-door neighbor in *Brunswick*, hailed her in a nominating speech as a "lady who by her experience and life qualifies herself to be a teammate, a partner, to our great governor."

(10) *Ray LaHood*, R-*Ill.*, one of the nine Republicans who voted against the resolution out of concern that it could result in more cuts in welfare, veteran and federal employee benefit programs.

3.4.3 Discussion

This relationship is often indicated by a person performing an activity in a location that requires that they live there. This includes working in a location, representing

a location in government, owning a home in a location, etc. There are a large variety of roles that are indicative of this example. Similar to the employer–employee relationship, these roles will need to be generalized in order to achieve high recall. There also appear to be many examples where a preposition is the primary indicator of the relationship: notable "in" and "of". Also this relationship sometimes occurs in addresses.

3.5 Organization–Location

3.5.1 Specification

1. Descriptions close to the location are valid: "north of", "south of", etc.

2. Airline destinations are indicative of this relationship (e.g. "AirlineX added a flight to Paris.")

3.5.2 Examples

The following are ten randomly chosen positive candidates of this relationship.

(1) When factors such as age, race, sex and education were taken into account, there was no significant difference in survival at six months between patients who chose aggressive treatment and those who sought so-called comfort care, reported rsearchers led by Dr. Jane C. Weeks of the *Dana Farber Cancer Institute* in *Boston*.

(2) Miss Morgan, the strictest of the strict at the *Academy of Ballet Florida* in *West Palm Beach*, Fla., will come forward and enthuse as only Miss Morgan can, "Girls that was beautiful."

(3) Nizar Hamzi, a professor of political science at *American University* in *Beirut* said, "I think the Israeli offer has to be viewed as a dramatic development."

(4) At *Children's Hospital and Medical Center* in *Seattle*, the cost of an ECG, an echocardiogram and related tests can cost up to $2,000, said cardiologist Dr. James French.

(5) A new, $40 million *Coca-Cola* bottling plant, one of four opened in Russia last year, stand on the edge of *Krasnoyarsk*, a regional capital in Siberia, more than 2,000 miles east of Moscow.

(6) Call (800)323-7373 or write AHI *International Corp.*, 701 Lee St., *Des Plaines*, Ill. 60016.

(7) (Susan Swartz is a columnist for The *Press Democrat* in Santa Rosa, *Calif.*)

(8) But the Texas court decision is of no comfort to the *Public Interest Research Group* chapter in *Ohio*, which sued the Buckeye Egg Farm of Cronton, Ohio, for consumer fraud for rewashing and repackaging eggs after the expiration data and redating them.

(9) "I was about 8, and I kept it inside the house, of course," says David McBride, 26, a chef with *Total Food Service Direction*, a *Miami* -based corporate dining company.

(10) Kings media relations assistant dies: Michael Jund, *Los Angeles Kings* media-relations assistant, died Sunday of a heart attack.

3.5.3 Discussion

In this relationship very few roles and/or keywords are used to indicate the relationhsip. Often the relationship is described by the word "in", "of", or "based". The juxtaposition of the two entities and the resulting argument structure is also often indicative of the relationship, especially in sports cases (e.g. "Los Angeles" "Kings"). The relationship also often occurs in addresses. Unlike the previous relationships, this relationship is not expressed using a role in most instances.

3.6 Lessons Learned

Each of the relationships discussed can vary a great deal in the manner in which they are expressed. The relationship can be expressed in a relatively standard way, or very differently from sentence to sentence. The idea of a "role" that one entity plays "for" another entity appears to very important in at least three of the relationships. These roles may be described by nouns, verbs, or larger phrases.

3.7 Important Elements of Relationship Detection

After studying several examples of these relationships in text, the following categories emerged as frequent elements involved in expressing a relationship (summarized in Table 3.1).

The first is the *basic pattern*. A basic pattern is underlined in the following sentence: "Mr. Smith is a director at ABC Studios". Each relationship has a large set of basic patterns, or words that occur between and around the two entities that indicate the relationship. A basic pattern is minimal in the sense that removal of any word from the pattern should not result in another basic pattern, or should no longer be indicative of the relationship.

Secondly, there are often *excess words* in the text besides the basic pattern. For example, in the following sentence: "Mr. Smith is a newly-hired director at ABC Studios". Here the description "newly-hired" can be removed to reveal the same basic pattern as in the example above.

Coreference resolution is critical to correctly identifying the entities involved in a relationship and achieving high recall. Coreference resolution includes identifying pronouns, appositives, and other text referring to a particular entity. The importance of coreference resolution can be seen in the following example. "Mr. Smith left ACME Inc., and he is now a director at ABC Studios."

Finally, relationship detection is often a result of *inference* from other knowledge. For example, in the sentence "Mr. Smith directed the film 'Lanterns', a production of

ABC Studios", the fact that Mr. Smith works for ABC Studios is indirectly indicated by two other relationships. Detecting relationships expressed in such a way requires inference rules. These rules can be very difficult to obtain, as well as difficult to agree upon. For instance, does directing a film for a particular studio indicate the employer–employee relationship? Answering such questions may require domain specific knowledge.

In order to understand the importance of each of the above elements statistics were collected indicating how often each is used. For each of the four relationships considered, counts of each expressive element were obtained on 50 examples. The results are presented in Table 3.2. These statistics help point out the differences between the realtionships studied.

Table 3.1: Elements of recognizing relationships and the analysis necessary to process this information are generalized from examples given in this chapter.

Element	**Type of Analysis**	**Examples**
Basic Pattern	Simple pattern matching around entities can detect the relationship.	Mr. Smith is a director at MGM Studios.
Excess	Perform language analysis to strip out unnecessary description to check for a BASIC PATTERN	Mr. Smith is a newly-hired director at MGM Studios.
Coreference	Coreference resolution is necessary to accurately detect entities involved in a relationship.	Mr. Smith left Miramax Studios, and he is now working for MGM Studios.
Inference	Knowledge needs to be retained and manipulated to derive other knowledge	Mr. Smith directed the film "Green Grass Lanterns", a production of MGM Studios.

Table 3.2: The number of occurrences of each expressive element in 50 relationship mentions.

	employer–employee	family	person–home	organization–location
BASIC PATTERNS	50	49	50	50
EXCESS WORDS	25	25	26	21
COREFERENCE	9	29	5	0
INFERENCE	3	7	1	0

Chapter 4

Data Processing

New York Times New Service text from 1998 was used for training and testing [11]. Several processing steps were then carried out to gather the necessary data for detecting relationships. These steps are shown in Figure 4-1.

The first step was to perform named entity detection on the data, using BBN's IdentiFinder [2], followed by sentence boundary detection. Sentence boundary detection was done be simply checking for periods, exclamation points, and of paragraphs that did not end a person's title. At this point only sentences containing at least one pair of the possible entity types were retained (e.g. (PERSON, ORGANIZATION) for the employer–employee relationship).

Part-of-speech tagging and sentence parsing were then performed on these target sentences in order to obtain the necessary linguistic information. The Brill tagger [3] and Collins parser [8] were used at this step. About 250,000 sentences that contained our target entities were processed.

For each relation we randomly sampled several thousand candidates, and tagged them as positive or negative instances by hand. The number of positive and negative tagged candidates for each relationship is shown in Table 4.1. Note that the different relationships have very different frequencies in the training data: in the training set around 29% of all person/company pairs involved a positive instance of

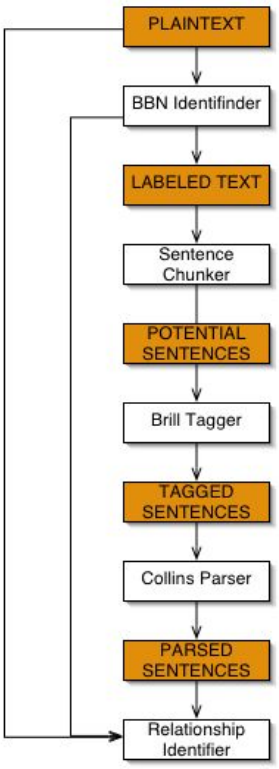

Figure 4-1: Diagram depicting data processing steps included named-entity detection, part-of-speech tagging. The intermediate states of the data are depicted in darker squares, while the processing steps are depicted in white squares.

the employer-employee relation, but only 2% of the person–person candidates were positive instances of the family relation. Because of the low percentage of positive candidates for the family relationship, only 3000 of the over 16,000 negative candidates in the training data were actually retained in the training set.

Table 4.1: The number of candidates of each relationship that were labeled positive and negative in the training and test sets.

	training set		test set	
	+	-	+	-
employer–employee	629	1671	145	355
family	400	3000	120	4590
person–home	280	840	117	383
organization–location	270	1206	93	407

Chapter 5

Features

Every pair of entities, e_1 and e_2, of the target classes that occur within a particular sentence, s is considered a "candidate", (e_1, e_2, s), of the relationship. For example, (Ross, Barry Goldwater, "Ross, 36, is Barry Goldwater's grandson."), is a candidate of the family relationship. Note that is an entity is mentioned twice in a sentence, each mention would be considered as part of a different candidate, and would not be considered together. This example will be developed throughout this chapter. Each candidate, x, is assigned a tag $y \in \{-1, 1\}$, depending upon whether or not the relationship actually holds for the candidate. Each candidate is characterized by a set of features derived from surface and linguistic information.

All the features from all the candidates are used to construct a feature vector ϕ for each candidate. These features are binary. They are assigned the value 1 or 0, depending upon whether an example contains that feature. This construction is captured by Equation 5.1.

$$\begin{aligned} \phi(x) &= \{a_1, a_2, a_3, ..., a_n\} \\ \text{where } n &= \text{ total number of features} \\ \text{and } a_i &= 1 \text{ if } x \text{ contains feature } i \text{ , 0 otherwise} \end{aligned} \tag{5.1}$$

Note: For the case of the family relationship, where both entities are of the same

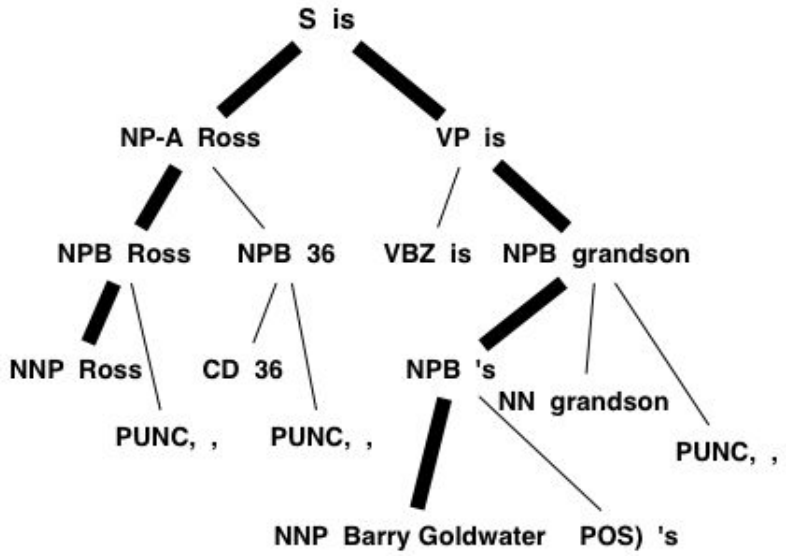

Figure 5-1: Parse tree of "Ross, 36, is Barry Goldwater's grandson." The path between the two entities is in bold.

class, e_1 is always the entity that appears first in the sentence.

5.1 Syntactic Features

The main syntactic features are based on the full syntactic parse of the sentence [8]. A parse of the sample sentence is depicted in Figure 5-1. The path through the parse tree between the target entities *(Ross, Barry Goldwater)* is bolded.

The full parse tree contains the part-of-speech tags of the words as well as the non-terminal tags of phrases of the sentence. It also contains the head word of each node in the tree. At the terminal nodes, the head word is simply the word from the sentence the node represents. At non-terminal nodes in the parse tree, "The head of a phrase is the element inside the phrase whose properties determine the distribution of that phrase, i.e. the environments in which it can occur." [22] Because of the way in which head words are chosen, often descriptions of entities will get pruned out at higher levels in the tree. For example, the head word of the phrase "the new mother" would be "mother". This can be usesful for generalizing relationship patterns around entities that may be described differently. More information on how a head word is

selected for a given phrase can be found in [18, 19].

Consider the path through the tree from Ross to Barry Goldwater. This path is characterized by the non-terminal/part-of-speech sequence " NNP NPB NP-A S VP NPB NPB NNP" and the head-word sequence "Ross Ross Ross is is grandson 's Barry Goldwater". These sequences are considered as separate features. In the head word sequence, the two entities in question will be replaced by their entity type, resulting in "PERSON is is grandson 's PERSON PERSON PERSON". Both sequences are also compacted so that duplicate sequential elements are removed. This leaves us with the following two features: "NNP NPB NP-A S VP NPB NNP" and "PERSON is grandson 's PERSON".

These features are designed for two purposes. The first is to increase the ability to recognize basic patterns involving words unseen in the training data. A single relationship may be expressed by a large number of basic patterns including words that do not occur in the training set. There may however be syntactic clues to the existence of the relationship, even if a pattern is unseen. For example consider the sentence, "Mr. Smith blank for ABC Studios", where blank is a verb. In many cases, no matter what the verb (runs, drives, directs, etc.), this construct will indicate the presence of an employer–employee relationship. The syntactic features may help generalize patterns so that each word of every pattern need not be known.

Secondly, these features may help eliminate excess words to better compare candidates. Considering only the path through the tree should help eliminate parts of the tree that are inconsequential to the existence of the relationship. In the example from Figure 5-1, the age of Ross is effectively "pruned" from our syntactic features by considering only the bold path.

Finally, compacting the syntactic features reduces the path to a form directly comparable with other paths where excess words have been eliminated. In the sample sentence, the appositive, `‘‘, 36,’’` is effectively pruned from the syntactic features. The syntactic features of the sentence of `‘‘Ross is Barry Goldwater’s grandson’’` would be the same as the sample sentence.

5.2 Between String

The full text string between the two entities is also included as a feature. This string can often contain a direct text pattern indicative of the relationship [4]. It may also contain text that occurs often in negative examples. Finally it will help in cases where tagging or parser errors occur, leading to incorrect syntactic features.

Unlike [4], no attempt is made to include text that occurs before or after the entity. The first reason for this is that, in order to be effective, the boundaries for these strings need to be generalized by several positive examples that are expressed in exactly the same way. We are not working with large amounts of data and our goal is to be able to detect relationships expressed in many ways. Secondly, the head word path features may draw in the necessary text from before and after the two entities. This can be seen in the sample where the word ‘‘`grandson`’’ is included in the syntactic features.

5.3 Ordering

The order in which the two entities appear in the text is also a feature (a binary flag indicating whether e_1 precedes or follows e_2). This feature is very important as the expression of a relationship can be very different depending on which entity is mentioned first. Consider the following two sentences.

```
He went to ABC Studios while in New York.
He spoke of New York while in ABC Studios.
```

In the first sentence, ABC Studios is located in New York, but this relationship does not hold in the second sentence. However the between string is the same in both cases. This ordering information will help to distinguish between such cases.

Note in the case of the family relationship, this feature is always the same as e_1 always precedes e_2 by definition.

5.4 Other Features

Several other surface and syntax-derived features are included that may have high correlation with either positive or negative examples. They are listed below.

- each word lying between the two entities
- the first word between the two entities
- the last word between the two entities
- bigrams of words between the two entities
- each word along the head word path
- the word at the top of the head word path
- bigrams on the head word path

5.5 Feature Representation of the Sample

The features of the sample pair *(Barry Goldwater, Ross)* are listed in Table 5.1.

5.6 Generalizing Features

One step that was taken towards generalizing features was to replace some entity mentions within the sentence with it entity type for all the features. For example, consider the following sentence along with it's generalized sentence.

`‘‘Hogwash,’’ said Jim Weldman, a spokesman for the National Federation of Independent Business in Washington, D.C.`

`‘‘Hogwash,’’ said PERSON, a spokesman for the ORGANIZATION in GPE, GPE.`

Consider the pair (Jim Weldman, D.C.) for the person–home relationship. In the features the appearance of “ORGANIZATION” instead of “National Federation of Independant Business” will help compare this candidate against other alike candidates.

This process of generalization was performed for all but the employer–employee relationship due to time constraints.

Table 5.1: Feature set of the sample pair *(Barry Goldwater, Ross)* from the sentence ‘‘`Ross, 36, is Barry Goldwater's grandson.`’’.

Feature Type	**Feature**
Part-of-Speech Path	NNP NPB NP-A S VP NPB NNP
Head Word Path	PERSON is grandson 's PERSON
Ordering	e_1,e_2
Word Between Entities	, 36 is
First Word Between Entities	,
Last Word Between Entities	is
Bigram Between Entities	, 36 36 , , is
Word Along Head Word Path	PERSON 's grandson is
Word at Top of Head Word Path	is
Bigram Along Head Word Path	PERSON is is grandson grandson 's 's PERSON

Chapter 6

Experiments

The goal of this project is to build a classifier that can use the features of each candidate to determine whether a particular relationship holds for that candidate or not. We applied a Support Vector Machine (SVM) classifier to the task, more specifically, we used SVM-Light [13] with a degree three polynomial kernel. (Tests on the development data showed that the polynomial kernel had better results than a linear kernel.)

SVM classifiers utilize feature representations of objects in order to classify them. This architecture is very flexible as features can be added or deleted as needed. SVM's use certain candidates, or support vectors, that are believed to lie near the boundary of the class in order to separate the two different classes. It chooses support vectors that allow for the greatest distance between opposing classes. For more information on SVM's see [9].

Along with the feature set as it was described in the previous chapter, and the classifier as it is described here, we conducted a number of other experiments designed to identify the good and bad points of the system. These experiments are described in the following sections.

6.1 Feature Noise Reduction Strategies

Experiments on development data show that by using all the features from all the candidates, we often aggregate many "noisy" features. These features may be outliers that occur infrequently. Also, the negative candidates do not belong to any well-defined class, and the features derived from them are not necessarily useful in this classification task.

Using these insights, four different noise cancellation strategies were identified. Classification for all four relationships was performed with each of these strategies.

(I) The default (base-line) approach: use all features from the training set

(II) Use features which occur with at least one positive candidate in the training set

(III) Use features which occur with at least two different candidates (positive or negative) in the training set

(IV) Combine both the second and third strategies

6.2 Sports vs. No Sports

Additionally, for the employer–employee relationship, experiments were carried out with and without data on the topic of sports. We manually tagged the training and test sentences which described sports events, and carried out experiments with and without this data. This followed an observation that sports data is stylistically quite different from most other genres within our dataset – we were interested to see if this particular genre had significantly different accuracy from other genres under our method. This experiment was not carried out for the other relationships because the differences in expression did not seem as important for these relationships.

6.3 Amount of Training Data

We also varied the amount of training data for each relationship to find how the effectiveness of the classifier changed with the amount of training data. Effectiveness of the classifier was measured by the highest F-score.

6.4 Effectiveness of Syntax Features

Finally the classifiers were also tested without the syntax features. This was done to help us determine how important the syntactic features were in the classifier. The syntactic features excluded were the following: the part-of-speech/non-terminal path from e_1 to e_2, the head word path from e_1 to e_2, the head word at the top of the path, each head word along the path, bigrams along the path.

Chapter 7

Results

7.1 Measuring Success

There are two main indicators of success measured in this thesis. The first is the maximum F-score. The SVM classifier assigns a score to each candidate it classifies and sets a threshold below which all scores are assigned to the negative class and above which all scores are assigned to the positive class. By varying this threshold, a variety of precision and recall values can be attained. The maximum F-score is the highest F-score achieved by varying this threshold.

The second measure of success will be referred to as the area-under-the-curve measure (AUC). It is intended to approximate the area under the precision versus recall curve. It is calculated by averaging the precision at recall = 10%, 20%, ..., 100%. At points where the precision is not known at the exact recall point, the precision will be interpolated from nearby points.

7.2 Results (F-scores)

The maximum F-score of the classifier is depicted for the experiments in the tables below. In Table 7.1, the maximum F-score of the classifier for each of the noise

reduction strategies and relationships is presented. The relationships are encoded as follows: EE = employer–employee, EE(NS) = employer–employee without sports data, F = family relationship, PH = person–home relationship, OL = organization–location relationship. A similar results table is presented for the case when syntax features were not used (Table 7.2).

Table 7.1: Maximum F-score of the classifier using each of the four features selection strategies. ([I] All features, [II] Positive features [III] Features that occur at least twice [IV] Positive features that occur at least twice)

Strategy	EE	EE(NS)	F	PH	OL
I	68.13	69.36	39.11	72.22	71.43
II	70.67	72.73	52.07	80.51	79.33
III	68.93	69.99	44.70	74.80	71.76
IV	70.42	71.67	52.73	79.11	79.04

Table 7.2: Maximum F-score of the classifier without using syntactic features.

Strategy	EE	EE(NS)	F	PH	OL
I	51.64	56.00	40.32	63.81	60.46
II	61.75	64.03	52.10	73.83	71.17
III	56.63	60.32	47.96	67.78	62.57
IV	62.51	64.08	54.30	71.96	70.51

7.3 Results (AUC)

The area-under-the-curve approximations are presented in Table 7.3 for each relationship and noise reduction strategy.

Table 7.3: Approximation of area under precision vs. recall curve calculate by averaging precision values at recall=10, 20, ... 100.

Strategy	EE	EE(NS)	F	PH	OL
I	67.28	66.96	26.82	76.19	73.13
II	74.13	75.15	43.85	84.73	81.80
III	72.18	71.50	36.47	79.14	75.33
IV	75.23	74.87	45.25	84.12	80.69

7.4 Results (Precision Versus Recall)

Precision vs. Recall graphs are presented for each relationship and noise reduction strategy in Figures 7-1, 7-2, 7-3.

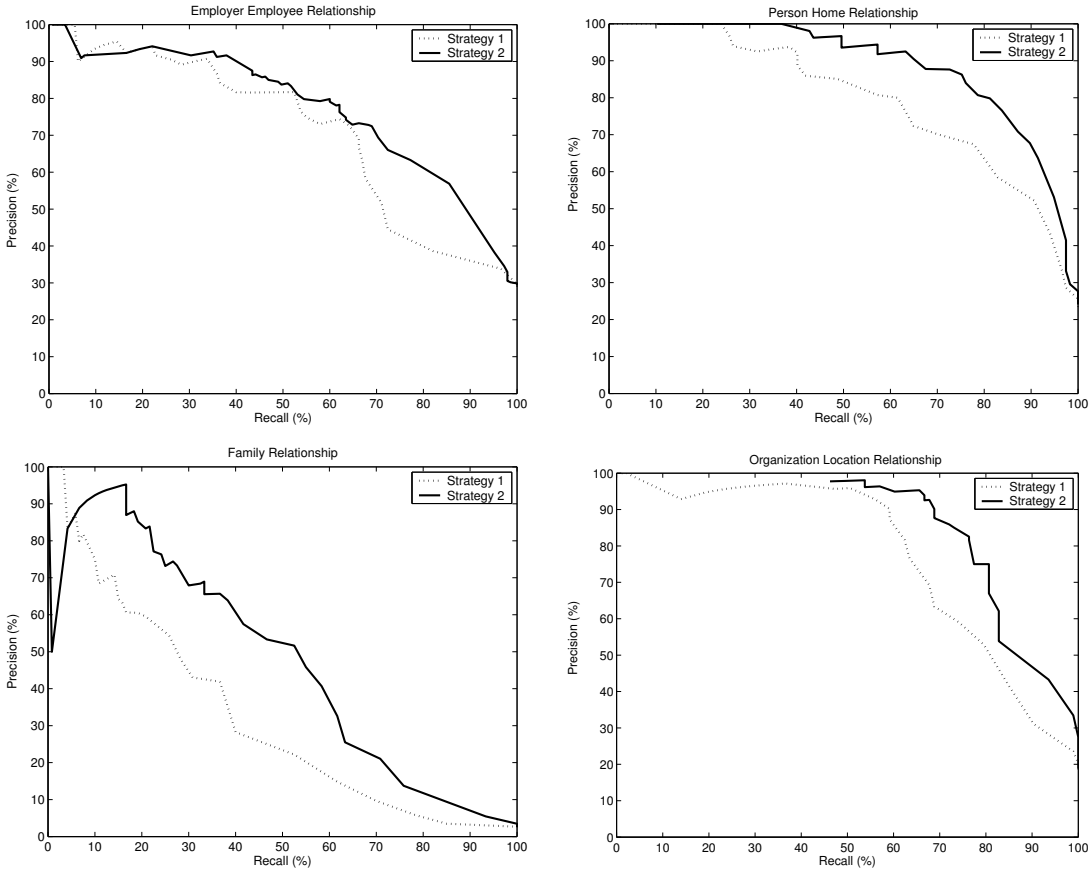

Figure 7-1: Precision vs. Recall graphs of all four relationships comparing noise reduction strategy I and II. Note that in the family relationship there is a large dip in the curve of strategy II at very low recall. This occurs because among the set of candidates tagged positively, only a few may actually be positive with a large amount negative. Then if positive candidate falls under the negative class as the threshold increases, without actual negative candidates also being added to the negative class, the precision will suddenly drop.

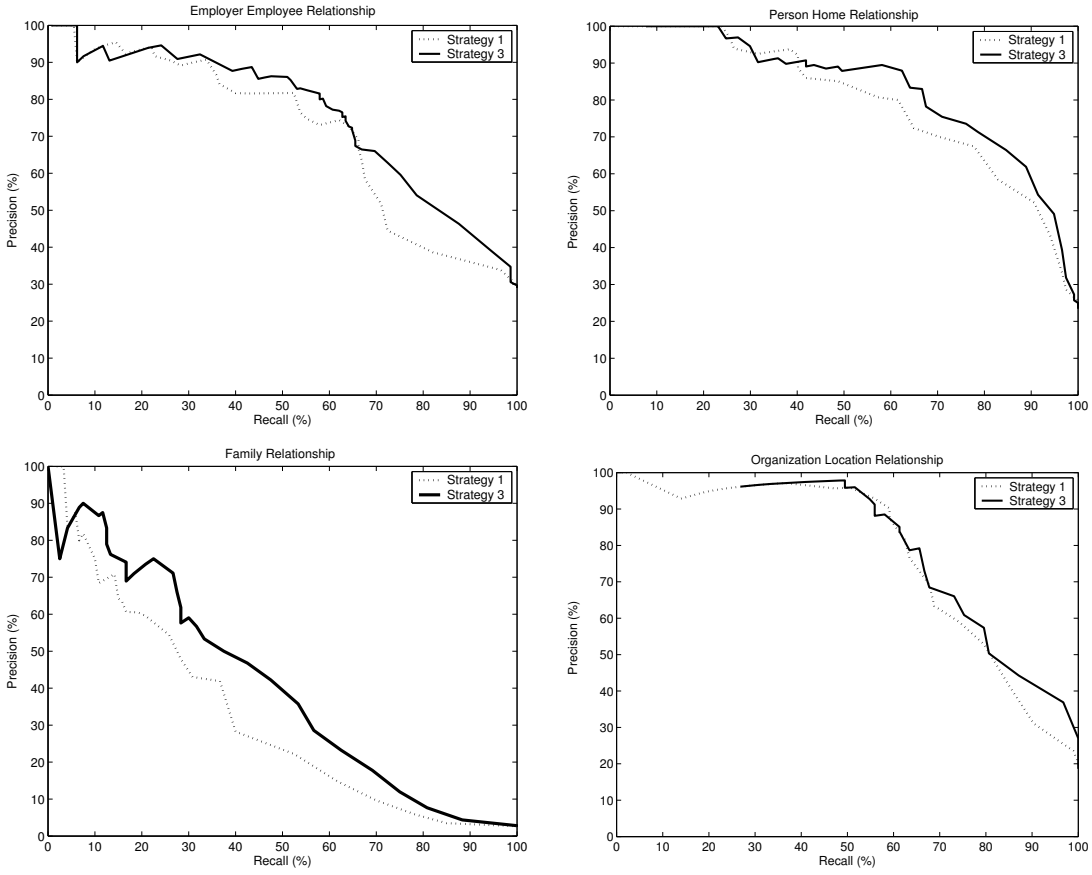

Figure 7-2: Precision vs. Recall graphs of all four relationships comparing noise reduction strategy I and III.

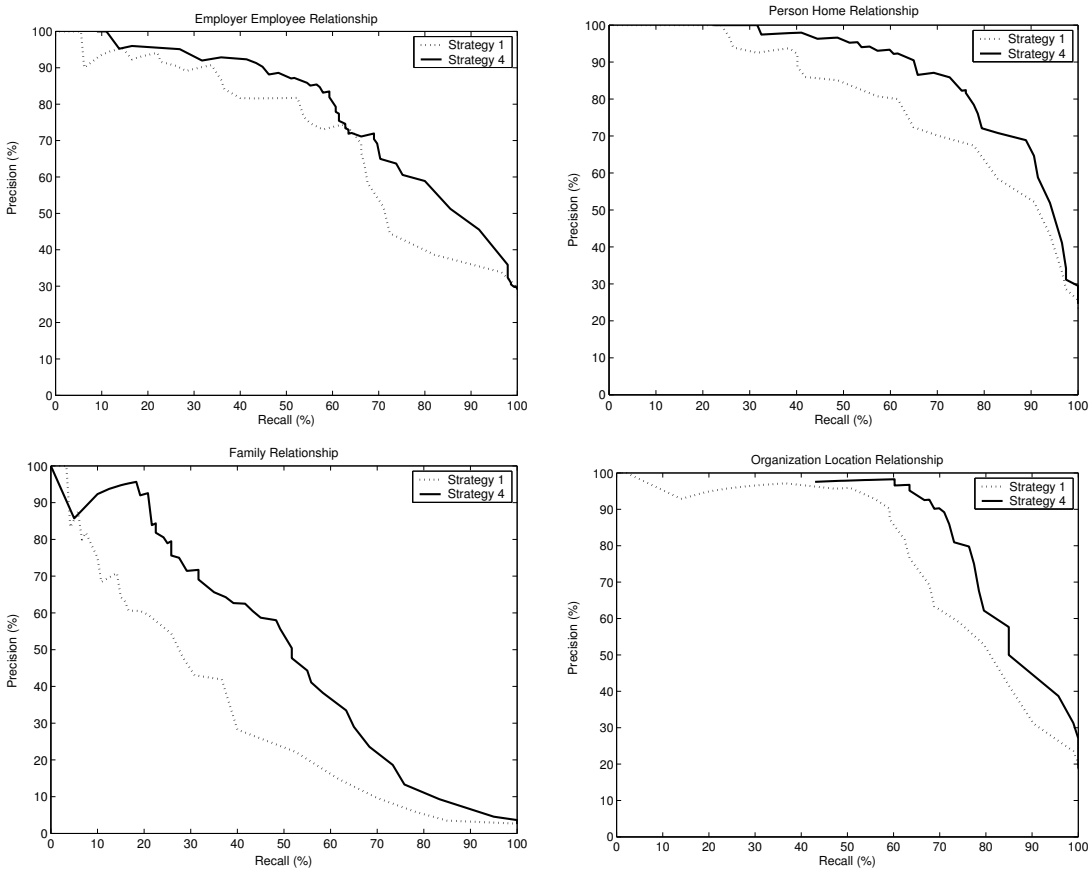

Figure 7-3: Precision vs. Recall graphs of all four relationships comparing noise reduction strategy I and IV.

7.5 Results (F-Score vs. Amount of Training Data)

The maximum F-Score was measured for various amounts of data for each relationship. The feature set used was that derived from strategy II. Graphs of the results are shown in Figure 7-4. The same data is plotted in Figure 7-5 with the training data plotted on a logarithmic scale.

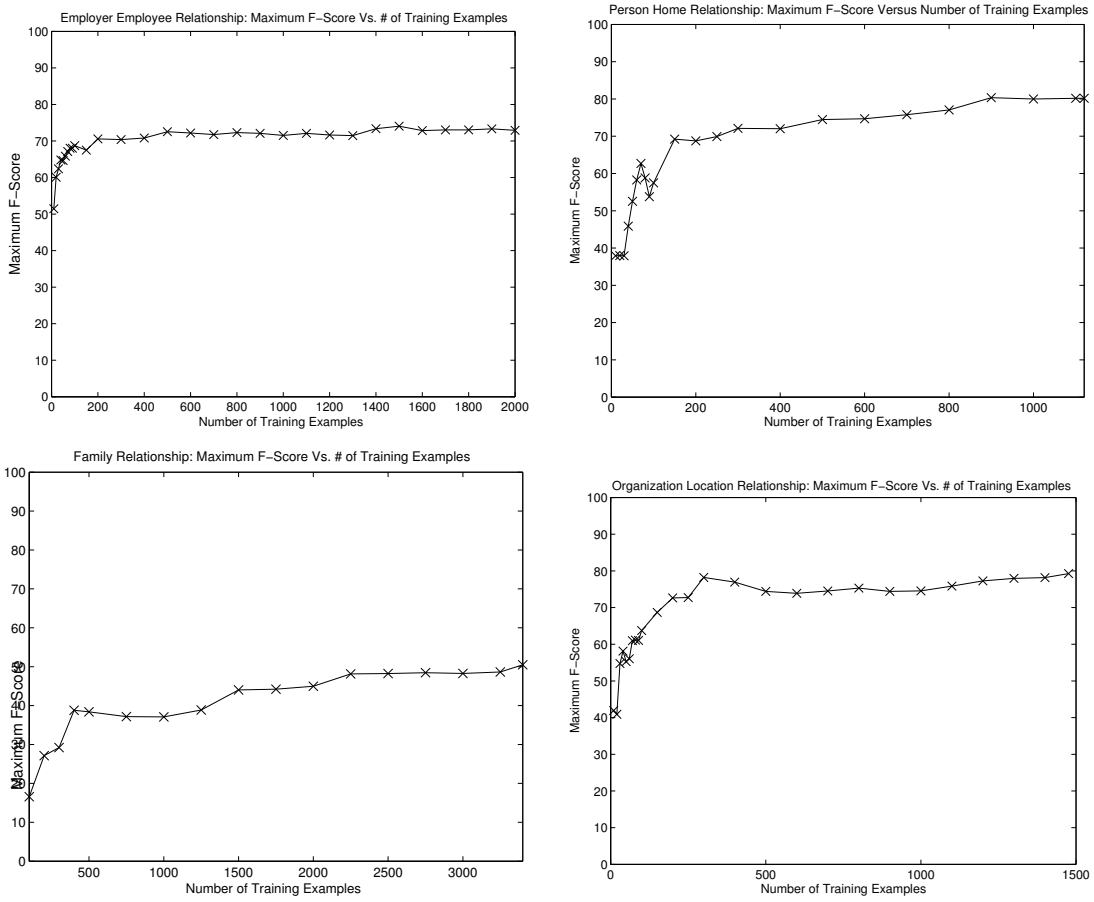

Figure 7-4: Maximum F-Score vs. number of training examples for each of the four relationships. The set of features was derived using strategy II.

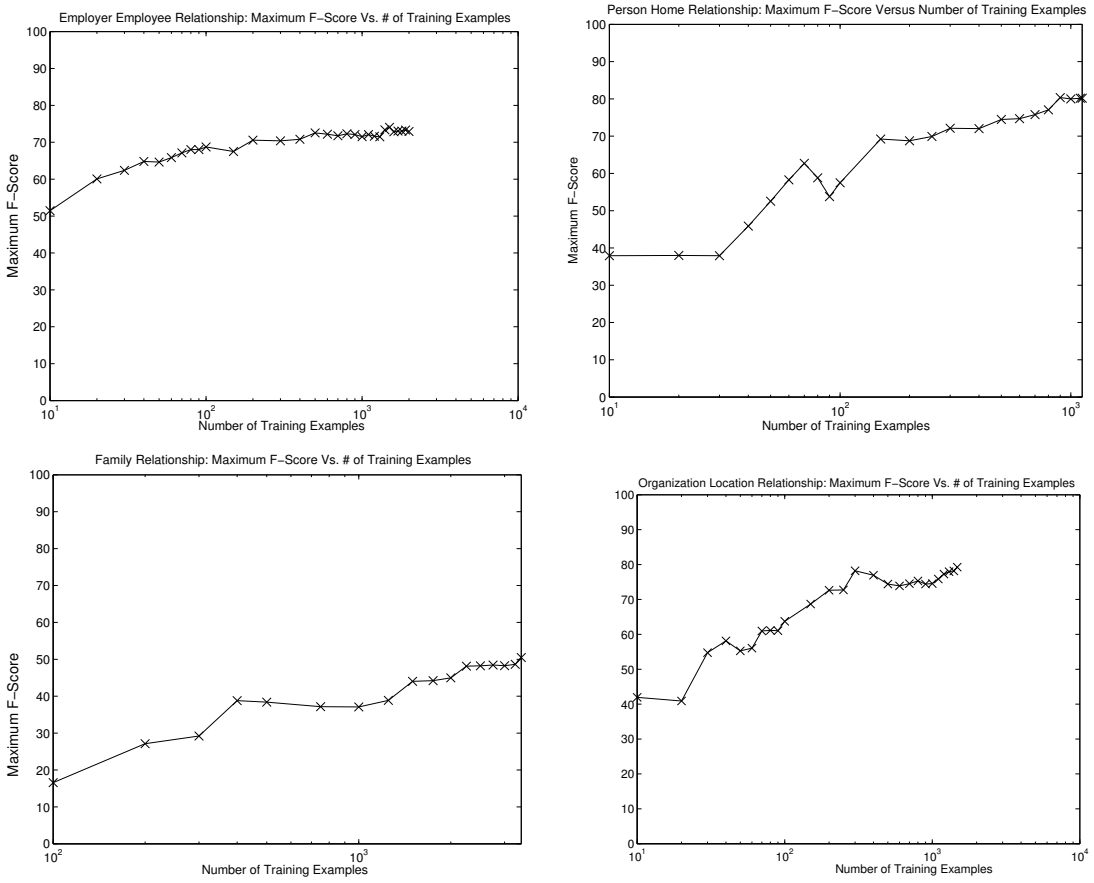

Figure 7-5: Maximum F-Score vs. number of training examples for each of the four relationships with the number of training examples on a logarithmic scale. The set of features was derived using strategy II.

Chapter 8

Discussion/Error Analysis

The results of the experiments identify some of the positive and negative aspects of this approach. First, several general aspects of the relationship extraction method will be discussed. Then, the results for each relationship will be discussed along with ways in which the system could be improved for each relationship.

8.1 Feature Noise Reduction Strategies

The use of features that occurred in at least one positive example (Strategy II) led to a significant improvement in all the relationships. For the person–home and organization–location relationships, this was a drastic improvement of about 8 points in the maximum F-score as well as the AUC measure. For the employer–employee relationship, the increase in the maximum F-score was not as high (2.54 points), but the increase to the AUC was 6.85 points. Similarly, for the family relationship, the maximum F-score was improved by 2.96 points, while the AUC improved by 17.03 points. The graphs in Figure 7-1 also display the effectiveness of this strategy. The precision vs. recall curves for Strategy II lie almost entirely above the curves for Strategy I for each of the relationships. These results provide strong support for the effectiveness of this strategy. They also support the hypothesis that negative examples do not consist of a well-defined class and will therefore contain very noisy

features. This strategy will be discussed further in the active learning section of the next chapter.

Using features that occurred in more than one training example led to smaller improvements to the maximum F-scores of the relationship classifiers: 0.80 to the employer–employee relationship, 5.59 to the family relationship, 2.58 to the person-home relationship, and 0.23 to the organization–location relationship. The improvements to the AUC were slightly higher at 4.90, 9.65, 2.95, and 2.20 respectivelly. Inspections of the graphs in Figure 7-2 also show significant gains while using this strategy. These gains are however less than those for Strategy II.

The use of Strategy IV led to improvements comparable to those of Strategy II. There was no significant difference between these two strategies across the relationships. However in the employer–employee and organization–location relationships, it appeared to perform slightly worse than Strategy II, but in the family and person–home relationships it performed slightly better.

8.1.1 Sports Data

Despite different results on development test sets, eliminating sports data from the training and test sets of employer–employee relationship did not significantly change the results. Inspection of the test set show that there were fewer "difficult" sports candidates than in the training set. An example of a "difficult" sports candidate is the following:

The Grizzlies made noise, trading Antonio Danials, the fourth pick last season, to the San Antonio Spurs for Carl Herrera and the rights to Felipe Lopez, the St. John's star and former schoolboy legend from Queens who was taken 24th by the Spurs. [11]

The method of expressing the relationship (having "rights to") is quite different than in other genres of text.

8.2 Syntactic Features

For almost every relationship and noise cancellation strategy, removing the syntactic features lead to a large decreases in the maximum F-score of the classifier. A notable exception to this was the family relationship, where removal of these features actually appeared to lead to slight improvement in the maximum F-score. However because of the generally low scores of this relationships, these results are not significant. Reasons for the failure of the syntactic features for this relationship will be discussed later in this chapter.

8.3 Amount of Training Data

The graphs from Figure 7-4 show the increase in maximum F-score with the number of training examples. The graphs show a steep rise in maximum F-score at the beginnings of the curve followed by slower steadier growth. The graphs in Figure 7-5 show the same data with the number of training examples along a logarithmic scale. These graphs show how the F-measure rises rapidly with small amounts of training data.

Generally the effectiveness of the classifier increases with the number of training examples and does not appear to taper off with the amount of training data shown. This suggests that additional training data would have helped improve the results of the classifiers. The employer–employee relationship seemed to have relatively little effect from increases in training data. The other three relationships had results that were increasing significantly with additional training data.

However, the amount of training data required for small gains in F-measure increases exponentially. This may be because many examples will be so close to previously seen data, they will make little difference on the classifier. However, it may be possible to be selective when adding training data by making sure that new examples will modify the classifier.

8.4 The Family Relationship

The relationship that performed the worst was the family relationship. From the statistics presented in Table 3.2, we can see that this relationship was expressed with coreference almost 60% of the time. This dependence on coreference is much higher than with the other two relationships. As there is no coreference tracking as part of the model, the system may have a difficult time with this relationship. Analysis of development test data shows that the recall rate on the examples containing coreference is only 28.8%, while the overall recall rate is 50.0%.

In some cases, the coreference is eliminated as "excess" by the syntactic features, allowing correct detection of the relationship. For example consider the following sentence.

`‘‘A California native, Orosco lives in San Diego with his wife, Leticia and three children’’` [11]. For this example the syntactic features are "NNP NPB NP NP-A PP VP S NPB NNP" and "PERSON lives with wife PERSON". While these features are not perfect indicators of the relationship, they may be close to similar examples in the training data. In order to solve this problem, a coreference model will need to be applied to the data.

Also, only about 2% of all the test data for this relationship is positive. This is very low compared with the other relationships. Because there is so much more negative data, it is more likely that negative examples get tagged positively (accounting for lower precision). Introducing a coreference model may be helpful in solving this problem because many of the positive training examples will generate less noisy features when the features are derived from the references to the entities which most directly indicate the relationship. For instance, in the above example, if the features were derived from the pair *(his, Leticia)*, they would be less noisy.

There are some other difficulties with detecting this relationship. One common way of indicating a family relationship is to utilize a common last name (e.g. "Edmund and Katie Hearst"). However most names separated by an "and" are not related in this way. Since we do not model first and last names, this phenomenon is difficult to

detect. Perhaps further modelling of entities (such as PERSON-FIRST, PERSON-LAST, or PERSON-FIRST-LAST), would help in these cases. A danger, however, is that the categories will be too specific to particular relations and general patterns will no longer be detected without a lot of training data.

8.5 The Employer–Employee Relationship

One of the big problems seen in the employer–employee relationship is poor generalization of the large set of words and phrases indicative of this relationship. A way to improve this would be to use clustering methods to detect words that fall under the same class. This might be done by introducing untagged data into the system, or by external clustering techniques.

There are also several false positives in the results of this relationship. Two-thirds of these can be attributed to mistakes by the named-entity tagger. This means that at least one of the entities that were part of the candidate, either the person or organization was not of the correct class. This is a serious problem. Ways to improve named-entity tagging include changing the tagger, tuning the tagger, or adjusting the relationship extraction system so that uncertainty information can be incorporated from the named-entity tagger. Other issues that lead to false-positives are parsing errors and the relationship *student-of* being marked as an *employer-employee* relationship, though these are relatively infrequent.

Lack of pronoun resolution accounts for a significant portion of the positive candidates marked negatively. However this is a much smaller problem than in the case of the family relationship.

Overall, the system performed rather well at extracting this relationship. In many cases it effectively eliminated excess description and is thus able to extract a large amount of information.

8.6 The Person–Home Relationship

The classifier for the person–home relationship performed fairly well when compared with the other relationships. Often this relationship was implied by having an affiliation with an organization in a certain location. As with the employer–employee relationship, the problem of generalizing the detection of roles people can have within organization that imply this relationship was difficult. Using word clustering may be helpful in this classifier as well.

Also, the number of false positives that can be attributed to mistakes of the named entity detector was high in this relationship, about 92%.

8.7 The Organization–Location Relationship

This extraction system performed well on this relationship. This may be largely due to the fact that this relationship is usually expressed in a small variety of ways ("-based", "in", "of" are the most common). Also, this relationship is usually straightforwardly expressed, with little use of coreference or description language.

There were relatively few false positives of this relationship, but mistakes of the named entity detector were present among this set.

Several of the positive candidates incorrectly tagged negative can be accounted for by parser errors. Especially difficult were cases where the relationship was expressed through an address.

8.8 Close Candidates

Another concern across all the relationships is the ability to discern between candidates within the same sentence. When considered one candidate in a sentence at a time, the implicit assumption is that the features of each candidate are independent. However, this is not the case. Consider the following sentence from the training set for the employer–employee relationship.

`It struck a deep vein among many gathered to celebrate Women In Film 's 25th anniversary and to honor the likes of producer Gale Anne Hurd , Columbia Tri-Star Motion Picture Group Vice Chairman LucyFisher , Women In Film founder Tichi Wilkerson Kassel _ and Streep herself .` [11]

The system will consider the possibilites of *(Streep, Women In Film)* and *(Tichi Wilkerson Kassel, Women in Film)*. While the syntactic features for these two pairs will be distinct, the other features may be similar enough so that both candidates are tagged positively. To prevent this there might need to be a way of discounting some features that are shared with other candidates, especially if only one candidate has strong syntactic features.

8.9 Near-Misses

Another occasional error that occurred was relationships close to the target relationship being tagged as positive. For example "student at" was often mistaken for instances of employer–employee, and "friend" was often mistaken for family. This may be because the syntactic features of these relationships are very similar. This demonstrates the difficulty of defining a relationship and is its specification. Often times these specifications can be very subjective, and do not necessarily reflect what the most "useful" information to extract would be. For example, would a user of the family relationship extractor, such as the intelligence community, mind learning the friends of a person along with the family? One way of avoiding this problem would be to extract the roles one entity plays for the other entity, and use a different system to group certain roles together to form a single relationship. Roles are generally well-defined in the candidates and are therefore not as subject to different interpretations.

8.10 Knowledge

A general observation that can be made is that knowledge of the world may be required to correctly interpret candidates. For instance, assume Mr. Smith lives in X-ville. In order for Ms. Doe to be his barber, she must live in X-ville. However, if she is his lawyer, it is not necessary that she live in X-ville. Experience tells me that I can hire a lawyer who lives anywhere, but it would be very difficult to hire a barber that lives far away from me. Making such distinctions is very easy for people who possess a great deal of knowledge of the world around them. However, there is no way for a machine to possess the same amount of knowledge (at least not yet).

8.11 Problems With Named-Entity Detector

Problems with the named entity detector account for a great deal of the false positives seen in the classification tasks. Problems with the named entity tagger may be causing noise in the training data as well. However, overall the use of the named-entity detector helps limit the amount of data that needs to be processed and achieves a precision of 93% [2]. The false positives caused by named entity detection mistakes are themselves evidence of the importance of named entity detection to the success of the this system. The system appears to rely strongly on the fact that the two entities are of the correct class. Methods for dealing with mistakes of the named entity detector will be discussed in the following chapter.

8.12 Measuring Success

One important note is that the success of the system is measured on candidates extracted using the named entity detector on single sentences. Given an entire document there would be a number of relationships that would be missed using this approach for two reasons. The first is that the named entity detector does not detect every entity correctly. Secondly there may be relationships expressed across more than one sentence. If the measure of success of the system was over a document and not over

our test sentences, the precision of the system would remain the same, however the recall of the system would be lower.

The goal of this project was to study sentence level relationship extraction. The next step would be to take the lessons learned from this project and apply them to document level relationship extraction. This will be discussed further in the next chapter.

Chapter 9

Conclusion and Future Work

9.1 Improvements to the Current Task

There are a number of improvements that can be made to the system. One of them is to test the features more rigorously to determine which features are useful, which need modification, and which features need to be added to better model the system.

Another important step is to introduce coreference resolution. This will help detect many relationships that are presently missed both in a single sentence, as well as across sentences.

Also, there may be a need to model entity categories in greater detail. For example, in the family relationship, there is a need to identify first and last names.

Using word clusters for generality may also be useful, especially for the employer–employee relationship. This could be done by identifying features that are generally positive indicators of the relationship, and then clustering features that often co-occur with these features. Also the use of WordNet [10], where words are grouped by lexicalized concepts may be useful.

Finally, it has been shown that increasing the amount of training data can be useful in these classification tasks. However the gains from additional training data grow very slowly, as most new training data if randomly selected, it similar to data

already seen. One way to remedy this is to carefully select new training dat using an active learning approach. This involves selected training data that have a high probability of impacting the classifier. This can be done by running the classifier over untagged data and selecting candidates that lie near the boundaries between the positive and negative classes. The classifier was least "certain" of how to classify these examples, and they therefore are likely to contain new information important to the classification task.

9.2 Reframing the Problem: Role-Discovery

One important lesson learned from this project is that relationship selection and specifications can be arbitrary, especially without a specific application in mind. Therefore it may make more sense to solve the problem of discovering the roles one entity plays for another entity. This problem is less subject to individual interpretation, and the data extracted can be used by a larger number of applications.

Often the role is expressed directly ("director", "sister", "base", etc.). In cases where a role is not directly expressed in the text, often an implied role can be deduced from the text. For example, in the text "Smith works for ACME Corportation" the role of "worker" is implied. In the text "Jagdish and Usha Singh have three children" the role of "spouse" is implied. The detection of both direct and implicit roles poses an extremely interesting problem.

9.3 Moving Toward Document-Level Relationship Extraction

This project has helped identify effective strategies for detecting relationships between entities expressed in a sentence. We have defined a feature set, studied how to reduce noise in the feature set, shown the importance of syntactic features, and identified many strategies that will help move us toward our goal of identifying how entities are related in text.

The next step in this study will be to develop a document-level relationship extraction system. The goal of this system could be to identify the same type of relationship discussed in this thesis, or to identify roles between entities. Studying the effectiveness of the individual features used in this thesis, especially the syntactic features, may provide clues for extracting roles. Detecting implicit roles may be more difficult, requiring specific rules or training data for each role.

This system should also address many of the issues discussed in this thesis in order to be effective. It should use coreference resolution in order to track each entity in the document. Also, the fallibility of the named entity detector should be considered. While the named entity detector can be very useful, it cannot be relied upon to detect all entities, or to be completely precise. The only way to detect every name or mention of an entity is to consider every sequence of words a possible entity.

Managing uncertainty will be critical to developing a document level relationship extraction system. Entity detection, part-of-speech tagging, coreference resolution, word clustering, and syntactic parsing all have some amount of fallibility and therefore some amount of uncertainty. It is important not to consider the output of any of these systems as 100% correct. One way of managing this uncertainty is to perform all these steps along with relationship detection concurrently. However, this is a very difficult endeavor. Further research into managing the uncertainty of each sequential step may prove to be fruitful.

9.4 Conclusion

This thesis has demonstrated the effectiveness of one relationship extraction strategy emphasizing syntactic information on a number of relationships. A number of critical issues have also been discovered and further work on these issues have been suggested. There are many ways of structuring the relationship extraction problem. I believe the most useful and interesting way of structuring this problem for future research is as a role-discovery problem over an entire document.

Bibliography

[1] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, 2000.

[2] Daniel M. Bikel, Richard L. Schwartz, and Ralph M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34(1-3):211–231, 1999.

[3] Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565, 1995.

[4] Sergey Brin. Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT'98*, 1998.

[5] Sergey Brin. Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conferen ce on Extending Database Technology, EDBT'98*, 1998.

[6] Mary Elaine Califf and Raymond J. Mooney. Bottom-up relational learning of pattern-matching rules for information extraction. *Journal of Machine Learning Research*, 4:177–210, 2003.

[7] Hai Leong Chieu, Hwee Tou Ng, and Yoong Keok Lee. Closing the gap: Learning-based information extraction rivaling knowledge-engineering methods. In *41st*

Annual Meeting of the Association for Computational Linguistics (ACL), pages 216–223, 2003.

[8] Michael Collins. Head-driven statistical models for natural language parsing. *Computational Linguistics*, 2003.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[10] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[11] David Graff. *The AQUAINT Corpus of English News Text*. 2002.

[12] Jerry R. Hobbs and David Israel. Fastus: An information extraction system. `http://www.ai.sri.com/natural-language/projects/fastus.html`.

[13] Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines*. PhD thesis, Cornell University, 2002.

[14] Elaine Marsh and Dennis Perzanowski. Muc-7 evaluation of ie technology: Overview of results.

[15] Scott Miller, Michael Crystal, Heidi Fox, Lance Ramshaw, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. Algorithms that learn to extract information–BBN: Description of the SIFT system as used for MUC. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[16] Scott Miller, Michael Crystal, Heidi Fox, Lance Ramshaw, Richard Schwartz, Rebecca Stone, Ralph Weischedel, and The Annotation Group. Algorithms that learn to extract information, bbn: Description of the sift system as used for muc-7. Technical report, BBN Technologies, 2000.

[17] Scott Miller, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. A novel use of statistical parsing to extract information from text. Technical report, BBN Technologies, 1998.

[18] Carl Pollard and Ivan A. Sag. *Information-Based Syntax and Sementics, Volume I Fundamentals*, volume 13 of *CSLI Lecture Notes*. CSLI Publications, Stanford, California, 1987.

[19] Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago, 1994.

[20] Ellen Riloff. Automatically generating extraction patterns from untagged text. In *Thirteenth National Conference on Artificial Intellligence (AAAI-96)*, pages 1044–1049, 1996.

[21] Dan Roth and Wen-tau Yih. Relational learning via propositional algorithms: An information extraction case study. In *IJCAI*, pages 1257–1263, 2001.

[22] Ivan A. Sag, Thomas Wasow, and Emily M. Bender. *Syntactic Theory: A Formal Introduction*, volume 152 of *CSLI Lecture Notes*. CSLI Publications, Stanford, California, 2 edition, 2003.

[23] Saic information extraction. `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html`, 2001.

[24] Stephen Soderland. Learning to extract text-based information from the world wide web. In *Knowledge Discovery and Data Mining*, 1997.

[25] Stephen Soderland. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1-3):233–272, 1999.

[26] Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. Unsupervised discovery of scenario-level patterns for information extraction. In *18th International Conference on Computational Linguistics (COLING-2000)*, 2000.

[27] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 71–78, Philadelphia, 2002. Association for Computational Linguistics.