Resolving *Other*-Anaphora

Natalia N. Modjeska

Doctor of Philosophy

Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh

2003

Abstract

Reference resolution is a major component of any natural language system. In the past 30 years significant progress has been made in coreference resolution. However, there is more anaphora in texts than coreference. I present a computational treatment of other-anaphora, i.e., referential noun phrases (NPs) with non-pronominal heads modified by “other” or “another”:

> [. . .] the move is designed to more accurately reflect the value of products and to put *steel* on more equal footing with **other commodities**.

Such NPs are anaphoric (i.e., they cannot be interpreted in isolation), with an antecedent that may occur in the previous discourse or the speaker’s and hearer’s mutual knowledge. For instance, in the example above, the NP “other commodities” refers to a set of commodities excluding steel, and it can be paraphrased as “commodities other than steel”.

Resolving such cases requires first identifying the correct antecedent(s) of the other-anaphors. This task is the major focus of this dissertation. Specifically, the dissertation achieves two goals. First, it describes a procedure by which antecedents of other-anaphors can be found, including constraints and preferences which narrow down the search. Second, it presents several symbolic, machine learning and hybrid resolution algorithms designed specifically for other-anaphora. All the algorithms have been implemented and tested on a corpus of examples from the *Wall Street Journal*.

The major results of this research are the following:

1. Grammatical salience plays a lesser role in resolving other-anaphors than in resolving pronominal anaphora. Algorithms that solely rely on grammatical features achieved worse results than algorithms that used semantic features as well.

2. Semantic knowledge (such as “steel is a commodity”) is crucial in resolving other-anaphors. Algorithms that operate solely on semantic features outperformed those that operate on grammatical knowledge.

3. The quality and relevance of the semantic knowledge base is important to success. WordNet proved insufficient as a source of semantic information for resolving other-anaphora. Algorithms that use the Web as a knowledge base achieved

better performance than those using WordNet, because the Web contains domain-specific and general world knowledge which is not available from WordNet.

4. But semantic information by itself is not sufficient to resolve other-anaphors, as it seems to overgenerate, leading to many false positives.

5. Although semantic information is more useful than grammatical information, only integration of semantic and grammatical knowledge sources can handle the full range of phenomena. The best results were obtained from a combination of semantic and grammatical resources.

6. A probabilistic framework is best at handling the full spectrum of features, both because it does not require commitment as to the order in which the features should be applied, and because it allows features to be treated as preferences, rather than as absolute constraints.

7. A full resolution procedure for other-anaphora requires both a probabilistic model and a set of informed heuristics and back-off procedures. Such a hybrid system achieved the best results so far on other-anaphora.

Acknowledgements

This dissertation would not have been possible without many people who provided support, encouragement, and inspiration.

First and foremost, I would like to thank my supervisor, Bonnie Webber, who has been a mentor, a role model, and source of inspiration. Bonnie has been patient with my relocating twice during the PhD program. Despite geographical distance, she has always been available to read drafts of conference papers and this dissertation, and to discuss ideas, hypotheses, and findings. Her firm and gentle encouragement, and constant support, have helped me to overcome obstacles and to grow as a researcher.

Graeme Hirst joined my dissertation committee in 2001, when I moved to Canada. He has been an excellent host and co-advisor, a source of ideas and assistance in various aspects of this work. Also, his copy-editing skills have been invaluable.

Katja Markert has been involved with this research almost from the beginning, and she joined the degree committee officially in 2002. Everything that I know about experimental design I learned from her. Katja has also been a source of ideas, practical advice, and moral support over the years.

During the second year of the PhD, I spent six months in Göteborg, Sweden. Elisabet Engdahl and Robin Cooper in the Department of Linguistics at Gothenburg University provided advice, support, and infrastructure during that time.

My physical environment since September 2001 has been the Interactive Media Lab in the Department of MIE at the University of Toronto. I would like to thank the Lab's director, Mark Chignell, for providing a desk and computer, and a part-time job, and also for creating a fun environment to work in.

Many thanks also to the faculty and students in the Computational Linguistics group of the Department of Computer Science at the University of Toronto, who made me feel welcome, and who shared their knowledge, time, tools, and resources, in particular, Diana Zaiu Inkpen, Eric Joanis, Melanie Baljko, and Suzanne Stevenson.

Over the years, many people have commented on the ideas in this dissertation and related work. I would like to thank Malvina Nissim, Ivana Kruijff-Korbayová, Paul Piwek, Hwee Tou Ng, and Fabrizio Sebastiani. I would also like to thank my colleagues and friends Jennifer Spenader and Sofia Gustafson-Capková, Melanie Baljko, Julia

Hockenmaier, and members of the Interactive Media Lab and the EPoCare project for excellent discussions and moral support through the more difficult parts of the PhD program.

Many people have shared their resources, software tools, time, and programming tricks. I would like to extend my deepest gratitude to Jason Rennie, Doug Rohde, Michael Strube, Christoph Müller, Joel Tetreault, Mirella Lapata, Diana Maynard, Aldebaro Klautau, Alexander Seewald, Julia Hockemaier, Richard Zemel, Daniel R. Allen, Jeff "japhy" Pinyan, and the guys at Toronto Perl Mongers.

Görel Sandström in the Department of General Linguistics at Umeå University in Sweden encouraged me to continue research activities and to apply for a PhD program outside Scandinavia. Without her, I would not be where I am now.

I also wish to thank my parents for years of support and encouragement. Their contribution cannot be quantified, but it has been substantial.

Finally, many thanks to my husband, David Modjeska, for financial and moral support, encouragement, inspiration, patience, and understanding.

This research was made possible through funding from the Engineering and Physical Sciences Research Council (grant GR/M75129) in the UK. Aspects of this research have been published previously as (Modjeska, 2002; Markert *et al.*, 2003; Modjeska *et al.*, 2003).

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Natalia N. Modjeska*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The phenomenon of other-anaphora

In the past 30 years significant progress has been made in the area of coreference resolution. Coreference resolution is concerned with identifying which pronouns, proper names, and definite NPs refer to the same object or individual, e.g.,

(1) In fact, it was Newman who encouraged Cruise to marry his then girlfriend, Mimi Rogers, which he did on 9 May 1987.[1]

Who married Mimi Rogers, Newman or Cruise? And whose girlfriend was she at that time? In other words, who do the pronouns "he" and "his" refer to? There exist several approaches to resolution of pronouns and definite descriptions (noun phrases with the definite article "the", e.g., "the President").

There is, however, more *anaphora* in texts than coreference. By anaphora, I mean a relation of dependence between two items in a discourse such that one of the items, the *anaphor*, is, in isolation incomplete and can only be properly interpreted by considering the meanings of the other item(s) in the relationship, *the antecedent(s)* (cf. (Carter, 1987; van Deemter and Kibble, 2000)). Consider, for instance, the following example,

[1] *The Scotsman*, 5 May 2001, p.12.

(2) Over four years, *Ukraine* would receive 75,000 million cubic meters of gas and 50,000,000–70,000,000 tons of oil, some of which would be passed on to **other European countries**.[2] (BNC)[3]

The expression that needs "assistance" with its interpretation in Example 2 is "other European countries". The phrase refers to a set of European countries *excluding* Ukraine, and it can be paraphrased as "European countries other than Ukraine". Note that (1) the two expressions do not corefer, since they refer to different entities: in one case, it is a country, in the other, it is a set of countries, all of them being situated in Europe; and (2) that the anaphor "other European countries" is barely interpretable without its antecedent "Ukraine".[4]

It is examples such as 2 above that this dissertation is about. I focus on *other-anaphors*, by which I mean referential noun phrases with non-pronominal heads modified by "other" or "another" and non-structural antecedents. (Throughout the dissertation I will be using the following terms to refer to this phenomenon: other-anaphora and other-anaphors. To refer to *all* NPs with the modifiers "other" or "another", including non-referential uses, which are explained below, I will use the term *other-NPs*.) What I mean by the definition above is as follows. First, an other-anaphor either refers to an entity in the speaker's and hearer's discourse model, or it has the potential to refer (following (Fraurud, 1992)), and so phrases with "other" and "another" that do not and can not refer (they are thus non-referential), e.g., idiomatic expressions "the other week" and "another day" and discourse connectives "on the other hand" and "in other words", will not be addressed. Also excluded from this dissertation are reciprocal phrases "each other" and "one another", elliptic constructions "one X ... the other(s)" and "one X ... another", and "one"-constructions "the other/another one". Second, there are examples of other-anaphors in which the antecedents are available

[2]The following notational conventions are used in this dissertation. In the examples, anaphors are rendered in **bold font**, antecedents are rendered in *italics*. With coreference chains, the whole coreference chain is marked.

[3]Examples in this dissertation come primarily from two sources: the British National Corpus (BNC) and the *Wall Street Journal* corpus (WSJ). The sources are explained in Sections 1.2 and 3.3.

[4]More precisely, it is the referent of the anaphor that is barely interpretable without taking into account the referent of the antecedent; see below for definitions of these terms. I will be using the terms "antecedent" and "anaphor" instead of "referent of the antecedent/anaphor" as they are less cumbersome.

structurally as well as anaphorically, e.g., in *list-constructions* such as Example 3 and *other-than constructions* such as Example 4:

(3) The finding probably will support those who argue that the U.S. should regulate the class of asbestos including crocidolite more stringently than the common kind of asbestos, chrysotile, found in most *schools* and **other buildings**, Dr. Talcott said. (WSJ)

(4) The Soviets, who normally have few **clients other than the state**, will get "exposure to a market system, " he says. (WSJ)

In list-constructions such as Example 3, the antecedent(s) appear within the same coordinated NP as the anaphor, to the left of a conjunction "and", "or", "but", "as well as", or "along with". Exceptions exist, e.g., the antecedent of the first "other" in Example 5 below is not "(hundreds of) people" that occurs in the list-construction, but "most dogs" in the beginning of the sentence:

(5) *Most dogs* live for about 10 years on average, and during their lives they will come into contact with possibly hundreds of people and **other dogs**, as well as other animals such as cats and horses. (BNC)

But such examples are rare. In the four years of studying other-anaphora, I have come across only three or four such examples.

In other-than constructions such as Example 4, the entity to be excluded from the scope of the anaphor (here "the state") directly follows the particle "than". In fact, I use other-than construction as a test for antecedenthood: if the anaphor can be paraphrased as an other-than construction with the antecedent directly following "than", then we have found the correct antecedent.

Because antecedents of other-anaphors in list- and other-than constructions are available structurally, a fairly unsophisticated procedure would suffice to retrieve them. In the examples I consider, an antecedent may occur anywhere in the previous discourse or the speaker's and hearer's mutual knowledge, and thus a search procedure must be initiated in order to find the antecedent.

With respect to their meaning, other-anaphors provide a *set-complement* to an entity already in the discourse model, the antecedent. They introduce a new referent, which can be an individual object ("the other dog") or a set of entities ("other dogs"), which do not need to be referenced individually. Both types of referents must, however, have something in common with their antecedents. For instance, Ukraine in Example 2 is a country and it is situated in Europe; schools Example 3, in one of the senses of the word, are buildings; and national states can be business clients (Example 4). I will not dwell on the semantics of other-anaphors, as I am not concerned with it; some previous work on "other" and "another", including their meaning, is summarized in Section 1.3. It is, however, important to understand some terminology I will be using and the discourse processing theory in which this work is grounded.

I distinguish three levels of representation that are necessary for an analysis of anaphoric expressions: a *discourse*, a *discourse model*, and a *knowledge base*. A discourse is a written text produced by a single writer or a transcript of a multi-participant conversation. It serves as a linguistic source for a discourse model constructed during the interpretation of the discourse. A discourse model is an information structure; it consists of a set of discourse referents and conditions associated with them, i.e., their properties and relations to each other (Webber, 1978). A discourse evokes a particular discourse model, with respect to which it is interpreted. The model grows/is updated as the discourse unfolds in time. A knowledge base is a collection of facts about the world — the objects, their properties, and relations. Since reference resolution is concerned with figuring out the meaning (sense and/or reference) of anaphoric expressions, I will be wandering back and forth between the levels of discourse and discourse model. A knowledge base is taken here to be an external repository of knowledge, primarily semantic in nature, such as WordNet lexical hierarchy (Fellbaum, 1998).

The term *antecedent* refers to a linguistic entity in a text. The corresponding entity in the discourse model is the *discourse referent*. An other-anaphor is interpreted with respect to the referent evoked by the antecedent (or antecedents, in cases with split antecedents). Many times, such a referent is available directly. However, there are examples of other-anaphors in which the entity to be excluded is *mediated*, by linking the antecedent to some other entity in the discourse model, the *anchor*. Consider, for instance, the following:

(6) She lifted *the receiver* as Myra darted to **the other phone** and, her mouth set in a straight line, dialed the number of Roman's office. (BNC)

In Example 6, the antecedent of "the other phone" is the NP "the receiver". However, paraphrasing the anaphor as "the phone other than the receiver" is infelicitous, even if intuitively correct, because the entity to which the anaphor provides the complement is not the receiver, but the telephone (say, "telephone_1"), of which the receiver is the part. This entity "telephone_1" may or may not have been mentioned in the discourse earlier, and, in fact, to interpret Example 6, it does not need to have been mentioned. Because we know that telephones normally have receivers, we can derive the anchor "telephone_1" from the referent of "the receiver", and then use it to interpret the anaphor.

I will be using the term "anchor" somewhat ambiguously. The first usage has just been illustrated with respect to Example 6. The second is synonymous with the discourse referent of the antecedent. While it might be desirable keep the two entities distinctly apart in some other contexts, it is not necessary here. This dissertation is concerned with finding an entity (or entities) with respect to which an anaphor is interpreted (what to exclude), whether this entity is directly realized in the text (e.g., Example 2), or whether it is implicitly given, related to some explicitly realized entity (e.g., Example 6). Thereby, I use the term "anchor" to refer to an entity or set of entities with respect to which an anaphor is *interpreted*. Diagrammatically this is represented in Figure 1.1.

This view of antecedent and anchor is different from that of, e.g., Vieira and Poesio (2000) who used the term "anchor" instead of the term "antecedent" in bridging examples such as Example 6. Vieira and Poesio would call the NP "the receiver" the anchor of "the other phone"; they reserve the term "antecedent" for coreferential cases, in which anaphor and antecedent refer to the same entity. With respect to other-anaphora, adopting Vieira and Poesio's terminology would mean that all antecedents of other-anaphors were called anchors (as other-anaphors and their antecedents are not coreferential), and a new term would have to be created for entities such as "telephone_1" in Example 6, with respect to which the anaphor is interpreted. I therefore use just the terms "anaphor", "antecedent", and "anchor". Also, in Chapter 4, in which I present

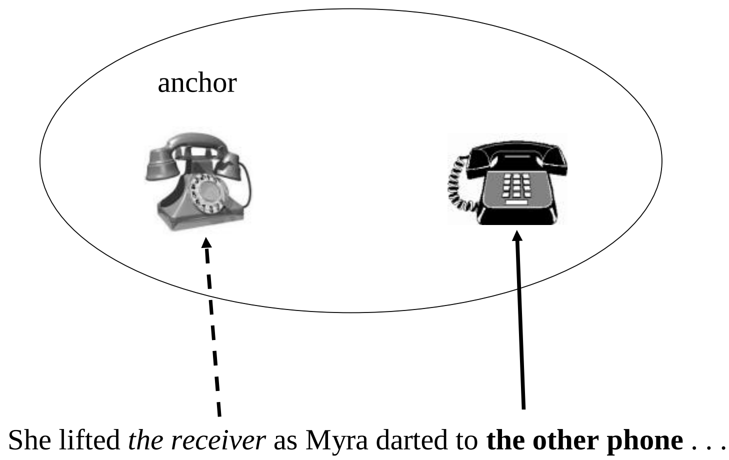

Figure 1.1: From text to discourse model: interpreting the anaphor and antecedent in Example 6.

a machine learning approach to resolution of antecedents of other-anaphors, I use the term "antecedent" ambiguously, to refer both to actual antecedents and to potential antecedents, i.e., all other NPs in a text besides the anaphor. Where it is necessary to distinguish the actual antecedent, I use more precise terms "correct antecedent" or "actual antecedent". Other terminology will be defined where necessary.

1.2 Why other-anaphora?

There are at least three good reasons to study other-anaphora. First, as mentioned in the previous section, anaphora is more than just coreference. Natural texts contain a variety of expressions that cannot be interpreted in isolation. Some of them have been studied previously, e.g., comparative adjectives (Staab, 1998) and alternative phrases (Bierner, 2000).[5] Others still await researchers' attention, e.g., modifiers

[5]This is a subjective list, for the purpose of illustration only; many more studies could have been included.

"such", "same", "similar", "different", and their cognates. Many of these words, including "other" and "another", are not as frequent as pronouns, proper names, or definite descriptions. (I am not aware of any quantitative study of this kind, though.) However, they are not as infrequent as they might seem at a first sight. For instance, in the British National Corpus (BNC)[6], a 100-million-word collection of samples of written and spoken language from a variety of genres, "other" and "another" belong to the top 200 most frequent words. "Other", tagged as adjective, is the 75th most common word. (There are also 35,164 occurrences of "other" tagged as noun, and 14,959 occurrences tagged as pronoun.) "Another" is the 159th most common word.[7] For comparison, demonstrative determiners "this", "that", "these" and "those" occupy the 22nd, 27th, 79th and 109th places respectively on the same list. Given the size of research literature on demonstrative pronouns and NPs with demonstrative determiners, e.g., (Kaplan, 1979; Linde, 1979; Gundel *et al.*, 1993; Asher, 1993; Sidner, 1983; Passonneau, 1993; Webber, 1991; Byron and Allen, 1998; Byron, 2002; Poesio and Modjeska, 2002), there is no reason not to study "other" and "another". Moreover, to fully understand what anaphora is about, all anaphoric phenomena must be addressed, including non-coreferential anaphors.

Second, other-anaphora interacts with other semantic processes, e.g., metonymies and bridging. Markert and Hahn (2002) pointed out that anaphora and metonymy resolution are often *co-dependent* and that metonymy resolution can benefit from anaphora resolution and vice versa. Consider their Example 7:

(7) We also tested *the printer Epson EPL-5600.* I liked **the laser**.

There are two benefits for metonymy resolution in this example. First, the information about possible anaphoric antecedents of "the laser" may help with choosing the correct metonymic interpretation from several possible readings, e.g., "laser" for "light". (A laser is an optical device and is one of the parts of a laser printer. In the example above, "the laser" refers to the same object as "the printer Epson EPL-5600".) Readings that do not allow for anaphoric interpretation would be dispreferred in Markert

[6] `http://www.hcu.ox.ac.uk/BNC/`

[7] These data are based on BNC frequency lists compiled by Adam Kilgarriff (`http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html`).

and Hahn's system. (The preference is for interpretations that establish anaphoric relations.) Second, no semantic constraints are violated in Example 7. Therefore, without the anaphoric information, a conventional metonymy resolution system based on selectional restrictions would not trigger the resolution process at all.

Metonymy (and metaphor) are not unusual with other-anaphors, as Example 8 illustrates:

(8) Columbia, a longtime Drexel client, won't provide current data on its junk. But its 17 big junk holdings at year end showed only *a few bonds that have been really battered*. *These* were *Allied Stores, Western Union Telegraph, Gillett Holdings, SCI Television and Texas Air,* though **many other bonds in Columbia's portfolio** also have lost value. (WSJ)

In Example 8, the anaphor "many other bonds" refers to debt certificates other than those issued by Allied Stores, Western Union Telegraph, Gillett Holdings, SCI Television and Texas Air. Company names frequently undergo `company-name-for-assets` metonymies; in fact, this metonymic pattern is very frequent in the *Wall Street Journal* corpus.

Often, resolving anaphoric references of other-anaphors requires reasoning. Consider, for instance, Example 9 below (and Example 6 in the previous section).

(9) Under the so-called Team Taurus approach, Mr. Veraldi and other Ford product planners sought the involvement of parts suppliers, assembly-line workers, auto designers and financial staff members from the initial stages of the development cycle.

The concept's goal was to eliminate bureaucracy and make Ford's product development more responsive to consumer demands. It was later applied to **other new-car programs**, including those that produced the Ford Thunderbird and Mercury Cougar. (WSJ)

In Example 9, the anchor of "other new-car programs" is the car program during which the Team Taurus approach was initially tested. There is no overt reference to that program, and its referent (and discourse anchor) must be *constructed* from the contextual information via some kind of bridging inference. Moreover, the construction of the referent *is triggered* entirely by the anaphor "other new-car programs" which cannot be correctly interpreted without it. (Alternatively, it has been suggested that the anchor is mediated by the phrase "the concept", as in "the concept's goal".) Likewise, in Example 6 in the previous section, a new referent "telephone_1" was introduced as part of the process of anaphor interpretation. Bridging and other inferential processes are not yet well understood. By studying what kinds of bridging and metonymic inferences are necessary to resolve other-anaphors, we can arrive at a better understanding of these phenomena as a whole.

Finally, a practical reason for studying other-anaphora springs from the demands of Natural Language Processing (NLP) applications. Any industrial system or research application that interprets or extracts information, finds intelligent answers to users' questions, or summarizes information, contains a Natural Language (NL) processing component. Within such a component, a reference resolution module is of crucial importance. Such a module keeps track of objects and individuals that the text or dialogue is about and their linguistic realizations, and resolves cases in which the intended referent of a phrase is unclear. Current reference resolution systems are only capable of resolving coreference relations. Consider, however, the following request to a virtual travel agent (adapted from (Jurafsky and Martin, 2000)):

(10) I'd like to order a return ticket from *Boston* to *San Francisco*, departing on December 5th, returning on December 12th. It's OK if it stops in **another city** along the way.

The virtual travel agent will have to figure out a lot of things in this request, including that "it" refers to the flight, and also that "another city" refers to any U.S. city on the way from Boston to San Francisco, excluding the origin and destination cities. A trivial piece of knowledge, this can be encoded as constraints ("can't stop in city X if the trip originates in X" and "can't stop in city Y if the trip ends in city Y"). Alternatively, this can be left for the reference resolution module to figure out. Note also that if the customer ordered a return trip to San Francisco, without explicitly mentioning the city of origin, the travel agent would still need to exclude Boston as a possible stop, relying on the extra-linguistic knowledge of where the conversation takes place in the physical or virtual space.

With the explosion of the World Wide Web and ubiquitous access to the Internet, applications such as information retrieval (IR), information extraction (IE) and question answering (QA) have become important means of finding information. The most common way to search for information is through a search engine such as Google or AltaVista. While doing a great job, present search engines still retrieve documents rather than answers to users' questions. And commercial search engines still use a combination of keywords and boolean operators, rather than a natural language such as English or Chinese. QA systems, unlike IR systems, attempt to provide a user with an actual answer rather than a collection of documents which may or may not contain the information the user is looking for. They take as input a single query, which can be formulated in a natural language and return a short passage of text (50 or 250 words). The future, however, lies with interactive systems that can accept multiple interrelated queries as in the example below:

(11) U: What company sells most greeting cards?
S: *Hallmark*.
U: How many?
S: 65 million cards per year.
U: How many do **other companies** sell?

Such a system must not only be able to retrieve an answer to each of the user's queries above, but also keep track of the user's questions and its own responses, to figure out that "other companies" refers to companies that sell greeting cards excluding Hallmark. Present QA and IR systems are not capable of handling other-anaphors, with the exception of the system presented in (Bierner, 2001), which, however, only handles "other" with structural antecedents such as other-than constructions "X(s) other than Y(s)" and list-constructions "X(s) and other Y(s)". (It also handles phrases "such as" and "besides".) It is clear that this is not sufficient to achieve the functionality required in Example 11.

1.3 Previous work on other-anaphora

Other-anaphors with non-structural antecedents have not been studied earlier, although a small body of literature exists on "other" and "another" and some of the constructions in which they appear. For instance, general grammar references such as (Quirk *et al.*, 1985) and (Huddleston and Pullum, 2002) provide guidance to how phrases with "other" and "another" are used and what they mean. The meaning of the modifier "other" was also briefly noted by Halliday and Hasan (1976), who classified it as a *general comparison item* that expresses difference between things, on a par with " different", "else" and "otherwise". The class of general comparison items further includes "same", "equal", and "identical" that express identity and "such", "similar", and "likewise" that express similarity. Halliday and Hasan noted that "other" has two meanings, "different" and "additional", leading sometimes to uncertainty of interpretation:

(12) I need some **other clothes**. — As well, or instead?

Halliday and Hasan did not discuss the meaning of "another". From the examples I have seen, it seems that "another" usually means "additional".

The semantics of "other" was discussed in detail by Bierner (2000), who recast it in terms of *alternative sets*, following (Rooth, 1992). As I am not concerned with the semantics of other-anaphors, but rather with how such phrases can be resolved in natural language applications (Section 1.4), I view their semantics as that of set-complementation, such that the anaphor provides a complement to an entity already in the discourse model, the antecedent. This allows me to treat "other" and "another" in a similar fashion, still allowing for flexibility of interpretation depending on the context of use.

List-constructions with "other" have been studied by Hearst (1992), who used them and the other patterns in Example 13 to acquire hyponyms from large corpora:

(13) a. $NP\{, NP\}* \{,\}$ *and|or other NP*

b. *such NP as* $\{NP,\}* \{(or|and)\}$ *NP*

c. $NP\{,\}$ *including* $\{NP,\}* \{or|and\}$ *NP*

d. $NP\{,\}$ *especially* $\{NP,\}*$ $\{or|and\}$ NP

and by Bierner (2000), who incorporated the analysis of list-other and other-than constructions in a natural language IR system. Other computational treatments of "other" include (Kamp and Reyle, 1993; Staab and Hahn, 1997; Staab, 1998; Salmon-Alt, 2001). Kamp and Reyle discuss how reciprocal "each other" can be resolved in the framework of their Discourse Representation Theory. Staab (1998) and Staab and Hahn (1997) presented a resolution procedure for "other" with omitted complements such as "one X ... other". (Salmon-Alt, 2001) is a corpus study of "autre" in French multi-modal dialogues within the framework of Cognitive Grammar.

This dissertation is the first to address other-anaphors from the perspective of how they can be resolved by a NL system. Specifically, I focus on *where and how to find antecedents* of other-anaphors, which is the first step towards their resolution.

1.4 The goal of this dissertation

The dissertation describes a procedure by which antecedents of other-anaphors can be found in naturally occurring texts. This procedure forms the backbone of a resolution method designed specifically for other-anaphora. While it might be thought that one resolution approach might fit all anaphors, there is evidence that suggests that different types of anaphoric phenomena respond differently to the same treatment (Strube *et al.*, 2002), and therefore they might require resolution algorithms tailored specifically for each anaphor type.

To constrain the search for antecedents, I examined a variety of syntactic, semantic and other constraints and preferences, e.g., antecedent surface form and syntactic function, their semantic class, and relation to the anaphor NP. To determine the extent to which these factors affect resolution of antecedents of other-anaphors, I designed and implemented two symbolic and several machine learning algorithms based on the Naive Bayes classifier. (Other machine learning methods were tested and rejected because of their poor performance.) All algorithms were tested on a corpus of examples from the *Wall Street Journal*.

As other-anaphors are a knowledge-intensive phenomenon, e.g., to resolve "other companies" in Example 11, the system needs to know that Hallmark is a company, I have also addressed the issue of knowledge acquisition from existing lexical resources such as WordNet and from the World Wide Web. (The work on the Web was done in collaboration with Katja Markert and Malvina Nissim (Markert *et al.*, 2003; Modjeska *et al.*, 2003).)

The final, hybrid, resolution procedure combines a Naive Bayes classifier with a set of informed heuristics, constraints, and fall-back procedures. This procedure achieved the best results to date on other-anaphora.

1.5 Thesis organization

Chapter 2 reports on a pilot study of other-anaphors in the British National Corpus. (Only NPs with the modifier "other" were considered.) Three hundred and fifty-eight occurrences of other-NPs were manually annotated along a variety of features, e.g., surface form of antecedent and anaphor, anaphor modification, distance, and type of lexical relation between anaphor and antecedent. A qualitative and quantitative analysis of the results of the annotation provided insights into what factors (syntactic, semantic, etc.) play a role in the interpretation of other-anaphors and therefore might also play a role in resolving their antecedents.

Insights from this exercise were used in designing two symbolic resolution algorithms, LEX and SAL, reported in Chapter 3. LEX is a heuristics-based algorithm; it resolves antecedents of other-anaphors on the basis of lexical information available from WordNet, pattern matching, recency, and class information for named entities. SAL is based on Centering Theory (Grosz *et al.*, 1995). It is an extension of Tetreault (2001)'s Left-Right-Centering, the state-of-the-art in pronoun resolution. SAL resolves antecedents of other-anaphors on the basis of grammatical salience and recency. (Grammatical salience here is taken to correlate with grammatical function of NPs.) The two algorithms were tested on a common corpus of 123 samples of other-anaphors (including "another") from the *Wall Street Journal*. LEX outperformed SAL by 32%, suggesting that lexical semantics is more useful in resolving antecedents of other-anaphors than grammatical salience. However, SAL resolved several samples of other-anaphors that involved metonymic, bridging, and pronominal antecedents, which LEX did not resolve (for obvious reasons — LEX was designed to handle only hyponymic antecedents and antecedents with the same head noun as the anaphor's). SAL was also more successful than LEX with samples of other-anaphors for which the necessary semantic knowledge was not found in WordNet. These two observations suggested that a combination of grammatical and semantic knowledge might lead to better resolution and possibly eliminate the need for developing a dedicated treatment for bridging and other less straightforward cases.

SAL used only one type of grammatical knowledge, grammatical function of the antecedent. There are, however, other grammatical factors that also might play a role in finding antecedents of other-anaphors, e.g., the linguistic form of anaphor and antecedent (e.g., proper name, definite NP), their gender, number, distance in words and sentences between anaphor and antecedent NPs, syntactic parallelism, etc. Chapter 4 examines these and other factors in detail, using a machine learning approach. At the core of the approach is the Naive Bayes classifier, which I show to be more suitable for resolving antecedents of other-anaphors than, e.g., a decision tree classifier. I present several Naive Bayes classifiers that differ with respect to how much and what kind of knowledge they use. One of the classifiers was trained on grammatical as well as semantic features (using knowledge from the WordNet lexical database). Another classifier, in addition to the features above, used semantic knowledge acquired from the Web by looking up lexico-syntactic patterns specific for other-anaphora and counting the frequencies with which they occurred. I also tested a variety of baseline classifiers, from those using grammatical information only, to those relying primarily on semantic knowledge (either from WordNet or from the Web), and a simple hand-crafted symbolic algorithm to indicate the difficulty of the task.

The best-performing classifier combined grammatical knowledge and Web-based semantics. While its performance was satisfactory, the resolution procedure it used was not yet a full decision procedure. For instance, the classifier did not take into account the fact that other-anaphors always require an antecedent. Also, a full resolution procedure must take into account other factors, e.g., syntactic constraints on antecedent realization (Chapter 3). Such a full resolution procedure is presented in Chapter 5.

Chapter 6 summarizes the research reported in this dissertation and suggests areas for future research. These areas include, e.g., (1) improving the learning framework and acquision of feature values; (2) extending the approach to anaphors with non-NP antecedents; (3) developing a special resolution mechanism for bridging and redescription cases; (4) further work on knowledge acquisition from the Web; and (5) embedding the hybrid approach in a real IE or QA system and testing it on other domains and languages.

Chapter 2

A Pilot study of other-anaphors in the BNC

2.1 Introduction

Any anaphor resolution algorithm takes at least three things and uses them to decide the intended antecedent of the anaphor. These are: (1) features of the anaphor; (2) features of potential antecedents; and (3) relations between the anaphor and potential antecedents (as well as, possibly, more general features of the text itself). For other-anaphors we do not have a priori any theory that tells us what features are relevant to resolving and, thereby, interpreting them. This chapter reports on an pilot study of other-anaphors in the British National Corpus, which identified several of such features.

Only NPs with the modifier "other" were considered. Three hundred and fifty-eight other-anaphors with full lexical heads were manually annotated along a variety of features, e.g., surface form of antecedent and anaphor, anaphor modification, distance, and type of lexical relation between anaphor and antecedent. This annotation exercise shed some light onto the nature of the phenomenon of other-anaphora and onto what resources and techniques are needed for their interpretation. The insights and statistical data produced by the study were subsequently used in the design of the LEX and SAL algorithms reported in Chapter 3. The corpus analysis also raised several theoretical and practical questions which will be addressed in future work, e.g., the role of modification in the interpretation of other-anaphors, as well as the type and amount of inferencing involved in their interpretation.

The chapter is organized as follows. Section 2.2 describes the data. Section 2.3 describes the annotation scheme which was developed specifically for this exercise. Subsequent Sections 2.4–2.7 spell out the details of the annotation and present and discuss the findings and their implications for the design of the resolution algorithm for other-anaphora.

This chapter extends work presented in (Modjeska, 2000).

2.2 The corpus

There are 185,308 occurrences of the word "other" in the British National Corpus. To make the analysis task manageable, I used a 1% random sample of the BNC. Random selection from the BNC is part of the Gsearch tool package (Corley *et al.*, 1999), which was used to extract the samples.

The 1% random sample of the BNC contained 1,820 occurrences of the word "other" in a variety of constructions, from referential other-NPs, to discourse markers, reciprocal "each other", idiomatic expressions, etc. I extracted all samples with "other", along with the context of eight sentences, i.e., the sentence with "other" plus four preceding sentences and three sentences that follow the sentence with "other". The word "another" was not part of the study. I annotated roughly 1/3 of this sample, or 445 other-phrases. Since the focus of this dissertation is on referential other-NPs, i.e., those NPs that have the potential to refer to individual objects (or sets of objects or eventualities) in the world, other-phrases that did not fulfill this condition, e.g., discourse markers with "other", idioms, etc., were dropped from further study. Another condition was that other-anaphors had full lexical heads; thus elliptic constructions, reciprocal constructions, and constructions with "one" were also ignored. Finally, constructions such as "X(s) other than Y(s)" provide antecedents by structural as well as anaphoric means. Other-than constructions were dropped from the study. List-other were considered in this study but dropped in subsequent work for two reasons. First, I will show Section 2.5, that list-other phrases seem to serve a discourse function different from that of non-list other-anaphors. Second, from a perspective of a resolution, list-other NPs are much easier to handle, as their antecedents are usually realized as the left conjuncts of other-NPs. (So far, I have seen only three or four examples where this does not hold.)

Of the 445 other-phrases in the annotated portion of the 1% random sample of the BNC, 358 samples of "other" were classified as potentially referring, having full lexical heads and non-structural antecedents. The remaining 87 cases were idiomatic and reciprocal other-NPs, etc.

2.3 Corpus annotation

2.3.1 The coding scheme

All phrases with "other" in the study corpus were annotated along a variety of features that had been identified as being relevant to resolving pronominal anaphora and some definite NP anaphora, e.g., antecedent NP type and its distance to the anaphor. Other features had not been previously used in annotating pronominal coreference or definite NP anaphora, e.g., all features related to NP modification. Finally, some known features were adjusted to reflect the nature of other-anaphors. One such feature is the type of relation between anaphor and antecedent.

I used the REFEREE coreference annotation scheme (DeCristofaro *et al.*, 1999) as the basis for designing the annotation scheme for other-anaphora and the REFEREE annotation tool for the annotation task. This tool offered the desired flexibility, was platform independent, and was available of-the-shelf.

The annotated features were grouped into features that pertained to the anaphor, features that pertained to the antecedent, and features that pertained to both the anaphor and antecedent (Figure 2.1). For all features, the first value is the default. Unless any other value was chosen by the annotator, the marked entity received the default value. This was necessary to preserve the completeness of the annotation scheme with respect to undesired examples of other-NPs such as, e.g., idiomatic other-phrases or other-than constructions (which were marked in the sample using the feature ANA IRRELEVANT,

ANA IRRELEVANT (no) (yes)
ANA PREOTHERMOD (none) (the) (THAT) (any) (EVERY) (SOME) (NUM) (P_NP) (P_pro) (no)
ANA POSTOTHERMOD (none) (NUM) (ADJ) (NUM&ADJ)
ANA OTHER RESTR (none) (RC) (PP)
ANTE FORM (no) (explicit) (implicit)
ANTE EXPLICIT (no) (DefNP) (IndNP) (Pro) (Dem) (PM) (P_NP) (P_PM) (P_Pro) (Adj) (P) (A) (DS)
ANTE IMPLICIT (no) (ComGround) (Infer) (Topic) (Undec)
ANTE CATAPHORIC (no) (yes)
ANA-ANTE DISTANCE (none) (0) (1) (2) (3) (4)
ANA-ANTE RELATION (none) (Parallel) (Identity) (Member) (Hyponymy) (Infer)
(Notes)

Figure 2.1: The annotation scheme for the BNC pilot study

but not annotated).

The anaphors were marked for the type of determiners, pre- and post-modifiers, e.g., quantifiers, demonstrative determiners, possessive modifiers, adjectives, ordinal and cardinal determinants, as well as relative clauses and preposition phrases. These are the features ANA PREOTHERMOD, ANA POSTOTHERMOD and ANA OTHER RESTR. They are explained in Section 2.3.2 and exemplified in Sections 2.4 and 2.7.

Discourse anchors of "other" can be evoked by an explicit linguistic expression, a piece of text, or they are extra-textual (situationally evoked). Implicit anchors received a tag "implicit"; explicitly realized anchors, i.e., those that have a linguistic antecedent, received a tag "explicit". Explicit antecedents were further classified according to their part of speech and/or syntactic category, e.g., definite or indefinite NP (DefNP and IndNP, respectively), pronoun (Pro), proper name (PM), demonstrative NP (Dem), adjective (Adj), possessive pronoun (P_Pro), possessive full NP (P_NP) or possessive proper name (P_PM), proposition (P), utterance (A), or discourse segment (DS). All antecedents were further marked for whether they preceded the anaphor, i.e. they were anaphoric, or whether they followed the anaphor, i.e., they were cataphoric, e.g.,

(14) If there is **no other system of air-conditioning**, and *the window* cannot be open all the time because of draughts or security, it should at least be opened once or twice a day for a spell.

Implicitly realized anchors, i.e., those that do not have explicit linguistic antecedents, were assigned a class depending on the degree of inference and type of knowledge involved. (Further details are in Section 2.4.)

While it might appear that it is always clear whether an anaphor has an implicit or explicit anchor, this is not the case, as familiar to anybody who has experience with corpus annotation. I resolved anchors on the basis of my linguistic intuitions, commonsense knowledge, and previous experience and training as a linguist. In this sense, the annotation should be considered preliminary. Where an example warrants a different interpretation from the one I propose, I briefly note that.

Distance between anaphors and antecedents was measured in sentence units, counting from the most recent mention. The values were "0" for list contexts "X(s) and other Y(s)"; "1" for antecedents in the same orthographic sentence as the anaphor; "2" for antecedents in the previous sentence; "3" for antecedents two sentences (or sentence fragments) away; and "4" for antecedents further afield. Implicit anchors and cataphoric antecedents, i.e., those that follow "other", instead of preceding it, received value "none" (distance not marked).

Discourse anchors that required inference, e.g., Example 6, were assigned a distance value only if some part of the text supported their derivation, as in the case in Example 6.

All anaphor-antecedent pairs were classified with respect to the type of lexical relation between them. More on relations and inference is in Sections 2.4 and 2.6.

The annotated items in the corpus were the anaphors, even though some of the features reflected properties of their antecedents and the relations between them. Antecedents were marked in the text and linked to their respective anaphors.

2.3.2 Annotation procedure

The annotation procedure was as follows.

1. ANA IRRELEVANT: If the other-phrase is a discourse connective, idiomatic expression, reciprocal or elliptic construction, construction with "one(s)", or other-than comparative, mark it as "yes" and go to next sample. Else mark it as "no" and go to Step 2.

2. ANA PREOTHERMOD: If the other-phrase contains a determiner/determinative, choose one: "the", "THAT", "any", "EVERY", "SOME", "NUM", "P_NP", "P_pro", or "no". Use value "THAT" for all demonstrative determiners; "EVERY" for all universal quantifiers; and "SOME" for all existential quantifiers. Possessive modifiers with full lexical heads, e.g., "John's" or "the department's", should be marked "P_NP"; pronominal modifiers, e.g., "his", should be marked "P_pro". Other-anaphors with cardinal determinants, e.g., "three other students", should be marked "NUM"; negated other-anaphors should be marked "no".

3. ANA POSTOTHERMOD: If the other-NP contains any further premodifiers besides "other" (they occur after "other"), e.g., "the other three students", mark them with respect to their type: "NUM", "ADJ", "NUM&ADJ".

4. ANA OTHER RESTR: Mark whether the other-anaphor is followed by a PP complement or relative clause.

5. ANTE FORM: Identify the antecedent of "other" by paraphrasing the anaphor as "X(s) other than Y(s)". For example, "dogs … other animals" can be paraphrased as "animals other than dogs". The phrase that follows the particle "than" is the anchor of the other-anaphor. If the anchor is realized as a linguistic expression in the text, e.g., an NP or clause, as opposed to the topic of the text or idea, mark it as "explicit". Else mark it as "implicit".

6. ANTE EXPLICIT: If the anaphor has an explicitly realized anchor, classify the antecedent with respect to its type of phrase/linguistic constituent, e.g., DefNP, IndNP, etc.

7. ANTE IMPLICIT: If the anchor of the other-anaphor is only indicated in the text and the text does not contain a single continuous expression which can be said to be related to the anaphor, classify the anchor with respect to its type. If the anchor is available via a common knowledge inference, e.g., "I hit him in the jaw. With my other hand I grabbed his throat.", mark it as "ComGround". If the anchor is the topic of the paragraph or section or any other discourse segment, mark it as "Topic". If the anchor must be recovered from the text via inferencing,

mark it as "Infer(ence)". If it is not clear what the anchor is or how to classify it, mark it as "Undec(idable)".

8. ANTE CATAPHORIC: If the antecedent follows the anaphor NP, mark it as cataphoric.

9. ANA-ANTE DISTANCE: Mark distance between anaphor and antecedent. See Section 2.3.1 for details.

10. ANA-ANTE RELATION: Mark the type of relation between anaphor and antecedents. See Section 2.6 for details.

11. Any notes and comments are typed in the field "Notes".

The data set was annotated by the author. The annotation has not been validated by other annotators, and no inter-annotator agreement score is available.

2.4 Types of anchors and their linguistic realizations

Anchors of other-anaphors can be evoked by a variety of linguistic expressions: definite or indefinite NPs[1], personal, demonstrative, or possessive pronouns, proper names, adjectives (Example 15), clauses (Example 16), or utterances (Example 17).[2]

(15) European Community officials were stoking fears last night of an all-out trade war with the United States after it was disclosed that *British* and **other European steelmakers** could face crippling new duties on exports to America.

(16) If *the patient is very heavy* or the carer cannot manage for **some other reason** [. . .]

(17) *How do I get my money back?* **Any other questions?**

[1] Indefinite NPs are NPs with determiners "a", "an", "some", bare NPs, and NPs with cardinal or ordinal numbers.

[2] The anchor in Example 16 is the proposition "the patient is very heavy.".

Multiple anchors ("split antecedent") are not unusual, e.g., Example 18, and can be viewed either extensionally — here, as a set containing all relevant individuals who attended the meeting, or intensionally, as "all the participants of the meeting mentioned so far".

(18) *Another speaker, Michael Traber, WACC's Director of Studies and Publications*, said: [...] In an address on "The Cultural Environment and Media Education", *Professor George Gerbner of the Annenberg School of Communication in Philadelphia* also stressed [...]. The participants voted to set up an Association for Communication and Theological Education to carry forward the discussions held at Yale. The Association will be coordinated by *WACC's former President, Dr William F Fore* [...] **Other participants at the meeting** included WACC's General Secretary, Rev Carlos A Valle, and John L Peterson, Chairperson of WACC's North American Regional Association.

Frequently, the anchor or part of it might need to be derived from a piece of text and perhaps some knowledge base (Examples 19 and 20, and also Example 6 from Section 1.1).

(19) *Bill Clinton* today prepares to stride across the political landscape as the world's most powerful man. But how does *he* shape up against *his* counterparts in **other countries**?

(20) Now the womens' task was packing the pilchards in the "bulks" in the Cellars, laid out and salted. After twenty-eight days they would be taken out, washed and packed in hogsheads, and pressed for about ten days. Once pressed, *the hogsheads, each weighting four and a quarter hundred-weight*, would be sold to the fish merchants for export, mainly to Italy and other Mediterranean countries. **Other pilchards** were kept for home consumption.

In Example 19, the anaphor is interpreted as "countries other than the U.S.", and the anchor "U.S." must be constructed from the antecedent "Bill Clinton" through a bridging (perhaps metonymic) interpretation process `leader-for-country` and knowledge of the political landscape of the world, specifically that Bill Clinton was (at the time the example was written) the president of the U.S. In Example 20, the linguistic antecedent of the anaphor “other pilchards” is “hogsheads, each weighting four and a quarter hundred-weight” (that were sold to fish merchants for export), while the anchor seems to be “pilchards that were sold for export” and it must be derived (via inference) from “hogsheads sold for export” to “hogsheads of pilchards sold for export” to “pilchards sold for export”. Information given earlier in the discourse that pilchards were salted, washed, and packed in hogsheads facilitates this inference.

It is not unusual for other-anaphors to make use of implicit anchors, relying on the speaker’s and hearer’s common cultural knowledge (Examples 19 and 21), general world knowledge (Example 22), knowledge of the utterance situation (Example 23), and/or ability to infer the anchor from what has been said so far (Example 20 and 24).

(21) The museums of the world are full of **other countries**’ art.[3]

(22) I slam a blow into his cheek, and it knocks his head against the wall. With **my other hand**, I grab his throat.

(23) Leo Jul 24–Aug 23 There have been **three other potential points of the year when much has been primed to change — not in the easiest of ways, nor in the most predictable of fashions –** and this is another one.

(24) When Suzie finally emerged she sent him back to bed. Now she and I sit in her little front room like middle-aged parents. [...] Suzie leans down to brush her lips against my cheek. Thanks again, she says. Then she stands up and crosses to **the other armchair**. She sits in it and drinks from a china cup.

Example 21 is an example of a bound anaphora interpretation: the entities to be excluded from the set of “other countries” are included in that set with each next iteration of the quantifier: for each country X, its museums are full of art from countries excluding X. This example relies on the common knowledge shared by the speaker and the hearer that many of the world’s countries (perhaps most) have art museums and that these museums usually collect and display art produced in the country of question as well as art produced in other countries. In Example 22, the anchor of “my other hand”

[3] *The Economist*, March 18, 2000, p.21.

is the implicit argument of the verb “slam a blow”. In Example 23, one might think that the phrase “Jul 24–Aug 23” is the time period to be contrasted with “other potential points of the year”. This is, however, incorrect; the phrase “Jul 24–Aug 23” is the time frame for the zodiac sign Leo. It is unclear from the example whether the horoscope was composed for “today”, “this week”, or “this month”. The reader of the horoscope knows, however, which one of these interpretations is the correct one. — Horoscopes usually span the inter-publication frequency of the periodical in which they appear. In Example 24, the reader is invited to create an anchor from two pieces of information: that the author of the text and Suzie are sitting on something (it is not clear what they are sitting on until we need to interpret the anaphor “the other armchair”) and that an armchair normally holds only one person (while a sofa or loveseat can sit two, three, or more people). In fact, in this example, the interpretation of the anaphor relies on something already known and, when interpreted, the anaphor adds new knowledge.

As Examples 20 and 23 showed, some anaphors require rather specialized knowledge. Likewise, Example 25 below relies on the reader’s knowledge of the process of distillation. Furthermore, the NP “some other liquor” is a *metaphor*, and so is its anchor. (I leave it to the reader to decide what it might be.)

(25) The move to the stage was a logical one for Eliot, so many of whose poems have dramatic qualities. In 1920, considering “the impotence of contemporary drama”, he had concluded that “The natural evolution, for us, would be to proceed in the direction indicated by Browning; to distill the dramatic essences, if we can, and infuse them into **some other liquor**”.

The anchor of “other” can be topic of a discourse segment:

(26) Well into the present century “Picklecock Alley” on Saltash Waterside was well supplied with sea-food shops. The Saltash Fair and pageant was a popular event of the 1930s, revived in the 1950s as Winkle Fair with “King Cockle”. The Mayoral party would receive a formal greeting to Waterside, “which was the proper old ancient Borough centuries before the town up and over was so much as dreamed of. Fish from the Water Tamar be pretty eating as all the world knows and we offer your worships this tribute according to ancient customs of

> our ancestors". Pollution largely ended the shellfish industry although there have been recent attempts to revive it. **Other ancient rights** have been eroded away.

In Example 26, there is no explicit antecedent and it is not clear what the anchor of "other ancient rights" might be: the rights for fishing, the custom of having a fair and pageant, both together, or something else? Nevertheless, it seems possible to interprete the anaphor in this example.

Table 2.1 summarizes frequencies of anchor types in the sample corpus. As evident from the table, the entity that other-anaphor excludes from the set is explicitly given in more than 3/4 of the cases (77%).[4] The majority of anchors are realized as NPs — 69%, including pronominal antecedents. Among them, full lexical NPs are clear leaders: definite NPs account for 23% of all anchors (including the implicit ones), indefinite NPs for 20%, and proper names for 19%. Nineteen percent of other-anaphors use implicit material as their argument: 10% require common knowledge anchors, 6% are inferred from the text, and 3% use discourse topic.

It is interesting that pronouns account for as little as 6% of all anchors. Psycholinguistic and computational linguistic studies have shown that pronouns are used to refer to the most salient objects. Since an other-anaphor provides a complement set to an entity already in the discourse, it is reasonable to assume that "other" with a pronominal antecedent would provide a complement set to an entity that is currently in the speaker's and hearer's center of attention. The low frequency of pronominal antecedents, compared with the frequency of other antecedent types, suggests that other-anaphors are able to access less salient discourse entities.[5] I will return to this issue in Chapter 3.

Demonstrative and adjectival antecedents are extremely rare in the BNC, with 0.5% frequencies each. Twenty-six percent of anaphors in the corpus use inference. Finally,

[4]This comprises anchors evoked by definite and indefinite NPs, proper names, pronouns, clauses (giving rise to a propositional anchor), utterances, discourse segments, demonstrative NPs, and adjectives: 276 samples over 358 total anchors in the study corpus.

[5]However, even if there are few pronouns in the study corpus, this doesn't mean that "other" isn't referring to the most salient or a highly salient entity. Both definite and indefinite NPs can be used for subsequent reference, and when reference is combined with predication, a definite NP is preferred over a pronoun. As Miller (1998) noted, "There is apparently a linguistic convention that accepts anaphoric nouns that are hypernyms of the antecedent." So, for instance, " a novel" can be subsequently referred to as "the book" as well as by "it": " I gave him a good novel, but the book/it bored him."

Type of anchor	Counts
definite NP	84 (23%)
indefinite NP	71 (20%)
demonstrative NP	2 (0.5%)
proper name NP	68 (19%)
pronominal NP	23 (6%)
adjective	2 (0.5%)
proposition	14 (4%)
utterance	8 (2%)
common knowledge	35 (10%)
inferred	20 (6%)
discourse topic	10 (3%)
discourse segment	4 (1%)
undecidable	17 (5%)
Total	358 (100%)

Table 2.1: Frequencies of anchor/antecedent types in the BNC pilot corpus.

in 5% of cases, it was impossible to determine the anchor.

These data suggested the following. When designing an algorithm for resolution of other-anaphors with full lexical heads and anaphoric antecedents, it would be reasonable to focus on NP antecedents, since they accounted for a good 2/3 of the total number of anchors in the corpus. Also, there was a good number of such antecedents in the corpus to allow for generalizations, while other types of antecedents were scarce and more data would be necessary to formulate procedures for resolving such cases.

2.5 Distance between anaphor and antecedent

The analysis of the distance data showed the following trends. From the perspective of reference resolution, it is important to know not only the syntactic type of antecedent, but also where it occurs in the text. Tables 2.2 and 2.3 show the distribution of explicitly realized anchors of "other" with respect to their proximity to the anaphor. In Table 2.2, the distance is calculated for each antecedent type; the values add to 100% crosswise, e.g, "24% definite NPs in list" means that of all definite NP antecedents, 24% were found in list-contexts. In Table 2.3, the focus is on list vs. non-list contexts. (The non-list value comprises "same", "previous", "two", "three" and more sentences away.) The values add up to 100% for each column, indicating how likely it is for the antecedent of "other" to be realized, e.g., in a non-list context as a definite NP. For example, for the configuration "definite" and "non-list", the probability is 0.33. For cases with multiple antecedents, the distance is given for the latest antecedent only. There were no cases in the study corpus in which it wasn't possible to resolve distance.

From these data, we can draw several conclusions. First, other-anaphora is a local phenomenon; the majority of (explicit) antecedents were found in the same sentence as other-anaphors or the preceding one (Table 2.4). These data might be compared with the data for pronominal anaphors.[6] Hobbs (1978), for instance, reported that 98% of pronoun antecedents in his corpus were found in the same sentence as the pronoun or the previous sentence. Second, as evident from Tables 2.2 and 2.3, there is a wide variation in the distribution of anchors of other-anaphors in list vs. non-list contexts.

[6]To my knowledge, no distance data are available for anaphoric definite NPs.

Distance						
Type of antecedent	list	same	previous	2 away	3 and more	not marked
definite NP	20 (24%)	35 (42%)	16 (19%)	2 (1.5%)	8 (10%)	4 (5%)
indefinite NP	28 (39%)	22 (31%)	14 (20%)	3 (4%)	1 (1%)	3 (4%)
proper name	25 (37%)	23 (34%)	12 (18%)	2 (3%)	6 (9%)	0
pronoun	3 (13%)	17 (77%)	2 (9%)	0	0	0
proposition	0	7 (50%)	6 (43%)	1 (7%)	0	0
utterance	0	2	3	0	3	0
discourse segment	0	1	3	0	0	0
demonstrative NP	1	0	1	0	0	0
adjective	1	1	0	0	0	0

Table 2.2: Distance between anaphor and antecedent in the BNC pilot corpus — variation within types of antecedents.

Type of antecedent	List	Non-list
definite NP	20 (26%)	65 (33%)
indefinite NP	28 (37%)	43 (22%)
proper name	25 (33%)	43 (22%)
pronoun	3 (4%)	19 (10%)
proposition	0	14 (7%)
utterance	0	8 (4%)
discourse segment	0	4 (2%)
demonstrative NP	1 (.5%)	1 (.5%)
adjective	1 (.5%)	1 (.5%)
total	78 (100%)	198 (100%)

Table 2.3: Distribution of antecedent types in list and non-list contexts.

Antecedent type	Percent
pronouns	100%
demonstrative NPs	100%
indefinite NPs	90%
proper names	89%
definite NPs	85%

Table 2.4: Proportions of NP antecedents in the BNC pilot corpus that appear within a two-sentence window.

This suggested that the two classes should be treated separately. More specifically, propositional, utterance, and discourse segment antecedents are rarely available in list-contexts "X(s) and other Y(s)", as it would require all elements of the list to have a similar type of denotation, which is rare. Furthermore, the percentage of indefinite antecedents in list contexts was significantly higher than that of definite antecedents (39% and 24%, respectively), and it was more likely for a list antecedent to be realized as an indefinite rather than a definite NP (37% and 26%, respectively). Also, many examples of list contexts with indefinite antecedents seemed to be of *generic* nature. Consider, for instance, Example 27:

(27) Film is capable of rising above the limitations of *language* and **other cultural barriers**.

Generic sentences express general statements about *kinds* of objects, rather than specific objects in the world. This seemed to suggest that other-anaphors may have different discourse roles in list and non-list contexts, and this is reflected in the distribution of its anchors. The exact discourse functions of other-anaphors in these different environments remain to be understood, but my preliminary hypothesis is that list-constructions are primarily used for their classifying properties, and that they do not introduce a new referent into the discourse model. (The classifying properties of other-anaphors are discussed in Section 2.6.) Non-list other-anaphors, on the other hand, do not only characterize the anchor as belonging to a certain class, but also *introduce new referents into the discourse* (a set referent in the case of "other Xs"). This was confirmed by an additional analysis of the data which showed that (set) referents of list other-anaphors ("other Xs" minus the anchor) were rarely referred to in the continuation of the discourse. The anchor and the set referent of an other-anaphor in list contexts are treated as a single set, and subsequent reference often means a reference to that set. For instance, in Example 28, the NPs "their" and "such victims" refer to the set of all victims of drugs, including the anchor "DES daughters".[7]

(28) *DES daughters* and **other victims of drugs** would be better off if their cases were taken out of the courts. Congress could create a compensation program to help such victims while protecting the national interest in encouraging new drugs.

Subsequent mentioning of the set-referent of an other-anaphor in non-list contexts, is however not uncommon, e.g., pronoun "they" in Example 29 below.

(29) Absorbed in doling out "Feeding Frenzy's" tidbits, the authors gloss over the root causes of Wedtech, namely the Section 8(A) federal program under whose auspices *the scandal* took place. They do at least come around to saying that the courts might want to end "rigid affirmative action programs." Programs like Section 8(A) are a little like leaving gold in the street and then expressing surprise when thieves walk by to scoop it up. **Numerous other scandals, among them the ones at HUD**, have the same characteristics as Wedtech. They take place in government programs that seem tailor-made for corruption.

Definite NPs were significantly more common in non-list contexts; 76% of all definite NP antecedents in the study corpus occurred outside the scope of a list other-anaphor. Furthermore, non-list antecedents were more likely to be of a definite NP type – 33%, compared with 22% for indefinites.

Proper name antecedents did not show a clear distribution pattern, and additional data collection and analysis would be necessary, though the data in Table 2.2 suggest a slight preference for list-contexts. When considering proper name antecedents from a more cognitive perspective, it is important to note that proper names usually have

[7]The following two examples are from the *Wall Street Journal* corpus (PennTreebank, release 2). No BNC examples were available at the moment of writing.

a special cognitive status and also that it is impossible to decide at the first glance whether the name is a first mention or it has been used earlier.

The third finding concerns definite and proper name NPs in non-list contexts. A surprising 10% of all definite NP and 9% of proper name antecedents were found three and more sentences away from the anaphor. In dynamic computational models of discourse, e.g., (Strube, 1998), the discourse model is updated with each new utterance (which many researchers consider to be a sentence). In models that adhere to Centering Theory (Grosz *et al.*, 1995), this often means that discourse entities evoked by a previous sentence that are not realized in the current sentence are dropped from the list of salient discourse entities at the end of the current sentence. The update procedure reflects shifts in the center of attention of the speaker and hearer — a coherent discourse centers around one entity. With respect to other-anaphors, the finding that 1/10 of all definite and proper name antecedents were most recently mentioned as far as three and more sentences away from the anaphor, suggested that other-anaphors can access referents that are no longer in the center of attention. This is perhaps not surprising, given the rich lexical content, which is available from the other-anaphor. Pronouns, for comparison, carry very little information and must rely on other mechanisms, such as salience.

In other theories of discourse processing, e.g., DRT (Kamp and Reyle, 1993), proper name NPs and definite NPs are treated differently, both with respect to each other and from how they are treated within, e.g., Centering Theory. In DRT, proper names introduce new discourse referents that are always accessible when interpreting subsequent sentences. Definite NPs, on the other hand, are generally treated as anaphoric, and they must either be *bound* to a referent that is already given in the discourse, or they are *accommodated*, i.e., added to the list of available referents. Accommodation is possible at three levels: locally, i.e., at the level of the current sentence, intermediately, and globally, i.e., in the so called main DRS (so that they become available for reference for all subsequently interpreted sentences, just like proper names do). These different possibilities for interpretation of proper names and definite NPs create a rich and complex structure, from which to pull the candidates for anchors of other-anaphors. I will not address this issue further; this topic is large enough for a separate

PhD project.

The distribution of antecedents of other-anaphors in terms of their distance from the anaphors had the following implications for the design of the resolution algorithm for other-anaphora. First, the majority of NP antecedents were found in a rather narrow window of two sentences. This means that a smaller window can be used in the resolution procedure (at a cost of not covering 100% of cases), and because of that, a smaller number of discourse entities would need to be considered when searching for the antecedent of an other-anaphor. Second, while not all left conjuncts of other-anaphors in list-contexts are their true antecedents, the majority are and therefore a rather simple lookup procedure can be used to resolve such cases. Cases with anaphoric antecedents require more complex resolution procedures.

2.6 Systematic lexical relations between anaphors and their anchors

Other-anaphors and their anchors are related to each other in two distinct ways. First, anchors of other-anaphors contribute to their compositional semantics; other-anaphors are uninterpretable without their anchors. A phrase “other Y(s)” always means “Ys excluding, or in addition to, some X(s)” where the “X(s)” are available from the preceding discourse, the utterance situation, or some other knowledge source. Second, other-anaphors trigger a presupposition that their anchors are of a certain kind. For example, in the phrase “dogs and other pets”, the presupposition is that dogs are (a kind of) pet. These presuppositions are licensed by the head noun(s) of other-anaphors and they restrict the set of potential antecedents. For instance, the referent of “chairs” cannot serve as anchor of “other pets” because chairs are not pets. Some of these presuppositions can be reduced to a rather small set of systematic lexico-semantic relations. Two such lexical relations have been observed by Bierner (2000) and Hearst (1992): the `instance-of` and `subclass-of` relations.[8]

[8] Hearst used patterns with list other-anaphors and other constructions to acquire hyponyms from a corpus. The `instance-of` and `subclass-of` relations are more general though, and hold as well for other-anaphors with anaphoric antecedents.

The `instance-of` (annotated as ANA-ANTE RELATION(Identity)[9]) is a relation between an individual object and a certain class to which it is said to belong, by virtue of being the anchor of the other-anaphor. In Example 30, for instance, the antecedent "Persia" is an `instance-of` the class of countries.

(30) *Persia* is exceptional in the number and variety of its weaving groups. **No other country** can boast the same range of masterworkshop, workshop, village and nomadic rugs [. . .]

A `subclass-of` relation (annotated as ANA-ANTE RELATION(Member) or (Hyponymy)) holds between objects representing concepts that stand in a hypernym–hyponym relation. For instance, in Example 31, a bow and arrow is identified as a kind of weapon.[10] Likewise, in Example 32, the anchor "the hall" is a kind (subclass) of object room.

(31) Every level has traps, baddies, bonuses and a huge nasty thing lying in wait at the end. Tiki, however, sports *a handy bow and arrow* and can also pick up **other weapons and handy methods of transport, such as balloons**, along the way.

(32) *The hall* is empty. There are lights in **the other rooms**.

Example 33 illustrates a special subcase of the `subclass-of` relation. A repeated form "benefits" is used to evoke both the anchor of "other benefits", invalidity benefits, and its class description:

(33) People on retirement pensions have to pay tax if they have any other source of income, so why shouldn't those who receive *invalidity benefit*? If they are very ill they can claim **other benefits, such as attendance allowance**.

From the perspective of resolution of antecedents of other-anaphors, there is no need to distinguish between the `instance-of` and `subclass-of` relations; they both roughly corresponds to an ISA relation in some lexical database. Examples such as 33 (annotated "ANA-ANTE RELATION(Parallel)"), on the other hand, allow to apply

[9]Lexical relations were annotated only for NP antecedents.
[10]It is not clear whether they are also identified as handy methods of transport.

string matching techniques. In subsequent chapters such examples will be referred to as "same-predicate" examples.

The third type of relation between other-anaphors and their antecedents — *re-description* — has not been described in the literature. Redescription is an *associative* relation. The class description evoked by an other-anaphor associates the anchor with a *different, but compatible,* class than the one to which it is known to belong. For instance, in Example 34, the British Clothing Industry Association, a trade organization, is identified as a sponsor of fashion shows ("sponsors other than the association, i.e., other than the British Closing Industry Association"):

(34) Until recently *the British Clothing Industry Association* subsidised the event, enabling Britain's designers to show their collections in an international venue. But *the association* has tired of being the sole supporter and **other sponsors** are needed.

Note that this information is also available from elsewhere in the text; the verb "subsidise" and the predicative NP "the sole supporter" predicate properties similar to that of the noun "sponsor". However, interpreting them would require a full semantic analysis of the sentence, while resolving the anaphoric references of "other sponsors" — and thus learning that the British Clothing Industry Association is a sponsor of fashion shows — is potentially a somewhat simpler task (though still a difficult one). The redescription relation is possible with common noun antecedents as well as with proper names:

(35) This enabled *the barley growers* to organise *themselves* effectively to protest to the authorities about *their* loss of land, and to challenge the monopolistic price-fixing of "middlemen". The book describes the experiences of **other oppressed groups in Mexico**, of outcast (Dalit) communities in India, and of fisherfolk fighting for their rights in the Philippines.

It should be pointed out that the redescription relation goes beyond a standard taxonomic classification and usually highlights some property of the object, its function or usage (Examples 36 and 37), or presents a speaker-specific point of view (Example 38).

(36) As of now, IBM Corp is offering an anti-virus service to its UK customers: this comprises *updates to anti-virus programs four times a year* — these provide users with "install-and-forget" automatic protection on MS-DOS, OS/2 and Microsoft Corp's Windows operating systems, and can detect viruses in the memory, on floppy and hard disks; if viruses are found in the memory, they are disabled, but if found on disks the user is given a recommended course of action on-line; **other services** consist of detection tools for any new viruses [. . .]

(37) In Wiltshire, *the working justices* consisted of one lawyer, Sir Robert Cherleton (later Chief Justice of the Common Pleas), and four local gentry, Nicholas Bonham, Sir Philip FitzWaryn, Sir Thomas Hungerford and William de Worston. All of these men gave **other service in the government**, as MPs, tax assessors, sheriff, commissioners of array and so on.

(38) The resulting report in 1960 listed professors' political activities, and said many had engaged in *"illicit love affairs, homosexuality, sexual perversion, excessive drinking* or **other instances of conduct reflecting mental instability**."[11]

One can test for whether an anaphor and its anchor are related through a redescription relation by the following test.[12] If the anchor can be paraphrased as "X is always a Y", then it is not a redescription relation, but one of hyponymy, e.g., "cats are (always) animals". A hyponymic relation expresses a profound, central characteristic of an object. Cats can also be pets, but they are first and foremost animals. If the anchor can be paraphrased as "an X can be seen as a Y" (and not all speakers of the language might agree with such an interpretation, e.g., that statistics are a form of factual abuse[13] or that taxes are a form of harassment), then the relation is likely to be of redescription. For instance, in Example 35, it is not the case that barley growers in Mexico have always been and will alway be one of the oppressed groups. This redescription describes a current state of affairs in Mexico (and it does not apply to other countries, as barley growers are not always and universally an oppressed group). Moreover, perhaps not even all Mexicans would agree with this characterization.

[11]Yahoo! News, the article "Feds Worked to Quash College Protests", published online June 9, 2002.

[12]I am grateful to Katja Markert for bringing this test to my attention.

[13]"Statistics and other factual abuse" was a title of an article in the 21 April 2001 issue of *The Economist*.

Redescription of the anchor may involve metonymic and metaphoric processes (Examples 39 and 40).

(39) When the dawn came, anxious viewers on the shore could see that the waves had taken with them the Eddystone lighthouse, *its eccentric architect* and **five other unfortunate souls**.

(40) *The human memory*, in common with **every other store**, has to be positively consulted before it will function.

In Example 39, the anaphor "five other unfortunate souls" refers to five people excluding the architect of the Eddystone lighthouse. (Note also that, although "the Eddystone lighthouse" is part of the coordinated phrase, it is not the antecedent of the other-anaphor.) Example 40 suggests an analogy between the human memory and some other types of storage.

Some examples of other-anaphors involve bridging inferences, e.g., Example 9 in Chapter 1 and Example 6 reprinted below as 41:

(41) She lifted *the receiver* as Myra darted to **the other phone** and, her mouth set in a straight line, dialed the number of Roman's office.

Redescription and bridging examples, and other examples that involved more than conventional `X-ISA-Y` inference received the tag "ANA-ANTE RELATION(Infer)". More work is necessary to identify the types of inference involved in the interpretation of such examples than what was said above; this is one of the areas of future research. With respect to straightforward relations such as `instance-of` and `subclass-of`, their systematicity is very attractive from the perspective of anaphora resolution, in particular for finding the anchors of other-anaphors. Specifically, the presupposition licensed by the head of the anaphor NP imposes a semantic constraint on its discourse anchor and thus restricts the set of what can be considered as antecedent. This constraint is the core method of the LEX resolution system, which is described and evaluated in Chapter 3.

Of the 358 samples of other-anaphors in the pilot corpus, 224 samples (62.57%) have NP antecedents. Of these 224 samples, a good two-third (68.75%) are samples

Relation	Annot tag	Counts
`instance-of`	identity	65
`subclass-of`	member	44
`subclass-of`	hyponymy	6
`subclass-of`	parallel	39
Subtotal		154
redescription, bridging, and inference	inference	70
Total		224

Table 2.5: Distribution of lexical relations among other-anaphors with NP antecedents in the BNC pilot corpus.

with `subclass-of` and `instance-of` relations between anaphors and antecedents (Table 2.5). The remaining 70 cases (31.25%) involve some type of semantic inference. These data suggested that in designing the resolution algorithm for other-anaphors, I should primarily focus on `subclass-of` and `instance-of` examples, at least to begin with. With an appropriate lexical database, such examples could be farely easy to resolve. Though as I show in Chapters 3 and 4, the quality of the lexical resource is of paramount importance. Furthermore, I show in Chapter 4 that using the Web as the source of semantic knowledge allows the algorithm to resolve not only `subclass-of` and `instance-of` examples, but also some redescription and bridging examples as well.

2.7 Modification and other-anaphora

Pre- and post-modifiers in other-anaphors seem to supply "additional" information that applies to both the anaphor and its antecedent. For instance, in Example 42, both "Harrison" and "Cornford", and the other scholars are said to be classical and anthropologically influenced.

(42) Eliot moves from such a primitive organization to discussing Greek drama, following the movement of *Harrison, Cornford,* and **the other anthropologically**

influenced classical scholars whom he had read.

Exceptions exist, however. In Example 43 below, the adjective "hemiplegic" is not applicable to the entity referred to by the antecedent, because hemiplegia is the paralysis of one side of the body, and therefore only one hand is affected.

(43) If he uses *one hand* on the cup handle, he should always have **the other hemiplegic hand** correctly positioned in front of him.

In speech, this difference is usually marked by a pause after the other-anaphor and a pitch accent on "hemiplegic". In writing, such additional information is often marked by commas, but examples lacking commas, similar to Example 43, are not unusual.

Similarly, in Example 44, the relative clause "who are engaged in the struggle for justice"[14] holds for both the anaphor and its antecedent:

(44) In his message of congratulations, WACC's General Secretary, Rev Carlos A Valle, wrote: "We welcome this award as a recognition of *your* courage and commitment in the field of human rights, and we trust that it will inspire **other groups and individuals who are engaged in the struggle for justice, both in Brazil and throughout Latin America**."

In Example 45, on the other hand, the relative clause "who could take part in debate but not vote" is exclusive of the antecedent, because "other members" are contrasted with "life peers with voting rights":

(45) The proposal got so far as a White Paper which suggested a two-tier system — *life peers with voting rights* and **other members who could take part in debate but not vote**.

In fact, in this example, the relative clause "who could take part in the debate but not vote" is a restrictive modifier and it is essential to complete the meaning of the anaphor.

When introducing the above examples, I used the word "additional", to describe the contribution of the modifiers to the resolution of other-anaphors. From the BNC data I had at my disposal, they did not seem to play an important role in finding antecedents

[14] I will not address the PP "both in Brazil and throughout Latin America", as its attachment is potentially ambiguous.

of other-anaphors. They seemed to describe less salient properties of the objects. For instance, to resolve the anaphor "the other hemiplegic hand" in Example 43, it is not necessary to know that one of the hands is hemiplegic, but it is absolutely necessary to know that the anaphor denotes a hand. Likewise, in the Example 45, it is not necessary to know that the other members of the Parlament could take part in the debate but not vote; it suffices to know that they are members of the Parlament.

Modification, in particular, processing of restrictive relative clauses, is an important component in the *interpretation* of other-anaphors, which is the second step in anaphora resolution. (The first one is identifying the correct antecedent.) Since the focus of this dissertation is on the first part of the resolution process, I did not analyze the study corpus further with respect to the issue of modification. (And this is a subject for future research.) Modification (in particular, pre-modification), however, turned out to play an important role in resolving some samples of other-anaphors — albeit from a different angle — when I tested the LEX and SAL algorithms described in the next chapter on a corpus of samples from the *Wall Street Journal*. I will return to the issue of (pre-)modification in Chapters 3 and 4.

2.8 Summary

In this chapter, I presented a pilot study into the phenomenon of other-anaphora on the basis of samples from the British National Corpus. The samples were annotated with a variety of features, e.g., type of antecedent, distance, and type of lexico-semantic relation between anaphor and antecedent (the latter for anaphors with NP antecedents only). The qualitative and quantitative analysis of the samples indicated what factors might play a role in the resolution of antecedents of other-anaphors.

For instance, the study showed that anchors of "other" can be evoked by a larger spectrum of expressions than previously noticed. Adjectives, clauses, utterances, discourse segments, as well as the utterance situation, can realize discourse anchors of "other". However, almost 2/3 of anchors of other-anaphors in the corpus were realized as NPs, and thus in designing the resolution algorithms in Chapters 3 and 4, I focused on resolving samples with NP antecedents. Second, the majority of NP antecedents were found in a rather narrow window of two sentences. Therefore, the LEX and SAL algorithms presented in the next chapter operate on a window of two sentences. (In subsequent work, Chapter 4, I used a window of 5 sentences.) Third, list other-anaphors were dropped from subsequent research; in the majority of list-other samples the antecedent is the left conjunct of the anaphor NP. To resolve such cases, a rather simple lookup procedure can suffice. Other-anaphors with non-structural antecedents, on the other hand, require a more sophisticated resolution procedure.

The BNC study corpus did not contain other-anaphors with the modifier "another". They were, however, addressed in subsequent research.

Finally, some other-anaphors stand in a systematic lexical relation with their antecedents. I gave examples of relations observed in the research literature and identified a new relation, redescription, which sometimes is licensed by the speaker's and hearer's common knowledge of the utterance situation and sometimes by their general world knowledge. (The type and nature of this knowledge is a topic for further research.)

I further showed that both explicitly given anchors of "other" and those that are mediated by the text or utterance situation might involve a variety of inferential processes, such as, e.g., bridging, metonymy, and metaphor. The precise nature and amount of inference needed to resolve such cases is a subject for further research. A good two-third of the anaphors with NP antecedents, however, trigger rather straightforward inferences encoded in lexical databases such as WordNet. The next chapter presents a symbolic algorithm LEX which resolves the anaphoric references of other-anaphors on the basis of information in WordNet and recency constraints.

Chapter 3

Two symbolic resolution algorithms for other-anaphora: LEX and SAL

In this chapter, I present two symbolic algorithms, LEX (for lexical) and SAL (for salience), to resolve antecedents of other-anaphors. (The reason for presenting two algorithms rather than one is given in Section 3.2.) LEX finds antecedents of other-anaphors on the basis of information in WordNet (Fellbaum, 1998), recency and syntactic constraints, and heuristics for Named Entity (NE) antecedents, presupposing that they have been classified into MUC-7 categories.[1] SAL is grounded in Centering Theory (Grosz *et al.*, 1995); it is an extension of Tetreault (2001)'s Left-Right-Centering (LRC), the best among state-of-the-art resolution algorithms for pronouns. SAL resolves antecedents of other-anaphors on the basis of their grammatical salience (which is correlated with the grammatical function of an NP).

Both LEX and SAL were informed by research into resolution of coreferring pronouns and definite NPs, and I will briefly review some of these approaches in Section 3.1 before introducing LEX and SAL in Sections 3.4 and 3.5 and their evaluation in Section 3.6. The algorithms were evaluated on a common corpus of examples from the *Wall Street Journal*. Data collection and preparation are described in Section 3.3.

In this chapter, I focus on *symbolic* resolution systems, informed by corpus studies, and on approaches which rely on the notions of focus/center of attention and/or

[1] For the current evaluation of LEX, NEs were manually annotated.

salience in tracking of anaphoric references. Machine learning approaches to coreference are reviewed in Chapter 4.

There are many other symbolic approaches to resolution of coreference and anaphora than the ones that are reviewed in this chapter; any theory of discourse processing must say something about how anaphoric items are addressed in that particular framework. They are, however, beyond the scope of this dissertation.

3.1 Symbolic approaches to coreference resolution

The task of coreference resolution is concerned with identifying which pronouns, proper names and full NPs refer to the same entity. In this section, I review three representative coreference resolution systems: Left-Right-Centering (LRC) (Tetreault, 2001), currently the state-of-the-art in pronoun resolution (for a good survey of pronoun resolution methods see chapter 18 in (Jurafsky and Martin, 2000)); a system for processing definite descriptions by Vieira and Poesio (2000); and the COCKTAIL system by Harabagiu and Maiorano (1999) which resolves both pronouns and definite NPs. These three systems have two things in common: (1) they were developed using evidence from corpora, and (2) they all share an assumption that salience of an entity in the discourse is correlated with its linguistic realization.

3.1.1 Pronoun resolution: LRC (Tetreault, 2001)

LRC is built upon Centering Theory's constraints and rules (Grosz *et al.*, 1995) as implemented by Brennan *et al.* (1987). Before presenting LRC, I will shortly introduce Centering Theory, as it is essential to understand both LRC and SAL.

Centering Theory is part of a larger theory of discourse processing developed by Grosz *et al.* (1995)[2]. The theory claims that discourse structure consists of three components: (1) a linguistic structure, which is a structure of utterances in the discourse; (2) an intentional structure, which reflects intentions and relations between discourse

[2]Various aspects of the theory were developed independently by several researchers, e.g., Sidner (1979); Grosz (1981); Joshi and Kuhn (1979); Joshi and Weinstein (1981). Grosz *et al.* (1983) integrated all previous work; the manuscript circulated since 1986 and was published in 1995.

segments, and (3) an attentional state, which models the speaker's and hearer's focus of attention at any given point in the discourse. The attentional state consists of two components: the local attentional state and the global attentional state. The local component reflects changes in the attentional state within a discourse segment; the global component models attentional state properties at the intersegmental level. Centering is concerned with local attentional state. Specifically, Grosz et al. claim that a speaker's choice of referring expressions affects the inference load placed on the hearer during discourse processing and the perceived coherence of utterances within a discourse segment. For instance, a reference by a pronoun indicates that the entity is currently in focus.

To model the speaker's and hearer's focus of attention at any given point in the discourse, Grosz et al. proposed the following. First, each utterance U_n in a discourse segment *DS* introduces a list of *forward-looking centers*, $Cf(U_n)$, and a single *backward-looking center*, $Cb(U_n)$. Forward looking centers roughly correspond to all discourse entities directly or indirectly realized in that utterance. The list of forward-looking centers is partially ordered to reflect their relative prominence in U_n. Ranking is based on a number of factors, from grammatical role to word order (especially fronting), clausal subordination and lexical semantics. In subsequent research, starting with Brennan *et al.* (1987), who operationalized and were the first to empirically test the claims of Centering Theory, *Cf* ranking was done by obliqueness of grammatical relation of the subcategorized functions of the main verb, i.e., *subject* > *object* > *object2* > *other subcategorized functions* > *adjuncts*. The first element in the list of forward-looking centers is the *preferred center*, $Cp(U_n)$. The backward-looking center is what the current utterance is about. Specifically it is the highest ranked element of the $Cf(U_{n-1})$ that is realized in U_n.

Second, to model the changes in the focus of attention, Grosz et al. proposed three types of transition relations between adjacent utterances: center continuation, center retaining, and center shifting. Brennan *et al.* (1987) proposed to further split center shifting into smooth shift and rough shift. The transitions relations are determined by the following criteria:

- whether or not the backward-looking center of the current utterance $Cb(U_n)$ is

	$Cb(U_n) = Cb(U_{n-1})$ or $Cb(U_{n-1})$ undefined	$Cb(U_n) \neq Cb(U_{n-1})$
$Cb(U_n) = Cp(U_n)$	Continue	Smooth shift
$Cb(U_n) \neq Cp(U_n)$	Retain	Rough shift

Table 3.1: Center transitions (Brennan *et al.*, 1987).

the same as the backward-looking center of the previous utterance *Cb*(U_{n-1}) and

- whether or not the backward-looking center of the current utterance *Cb*(U_n) is its preferred center, *Cp*(U_n) (Table 3.1).

In addition, there are the following constraints and rules (following (Brennan *et al.*, 1987)): For each utterance U_n in a discourse segment *DS*:

- There is precisely one *Cb*;
- Every element of *Cf*(U_n) must be realized in U_n;
- The backward-looking center *Cb*(U_n) is the highest-ranked element of $Cf(U_{n-1})$ that is realized in U_n;
- If some element of *Cf*(U_{n-1}) is realized as a pronoun in U_n, then the *Cb*(U_n) must also be realized as a pronoun;
- Transition states are ordered such that center continuations are preferred over center retains which are preferred over center shiftings.

Given these constraint and rules, Brennan *et al.* (1987) proposed the following pronoun resolution algorithm:

1. For each sentence, generate *Cb* and *Cf* lists;
2. Generate all possible *Cb*–*Cf* combinations;
3. Filter combinations by binding constraints and centering rules;
4. Rank all remaining combinations by transitions;

5. Select the highest ranking assignment.

Brennan et al.'s implementation of Centering Theory was subject to criticism. First, they made no provision for incremental resolution of pronouns, while it is well established that humans process utterances one word at a time. Second, the role of the second rule of Centering Theory, which ranks transition states (continuations are preferred over retains which in turn are preferred over center shiftings), has not yet been validated in pronoun resolution (Kehler, 1997a). Furthermore, this rule prevents incremental application of the algorithm. Tetreault (2001)'s LRC algorithm addressed these issues by making the algorithm incremental and by dispensing with the second rule of Centering Theory. The LRC algorithm first searches the current sentence, and if no antecedent for a pronoun has been found, then it searches the previous sentences, in the left-to-right order. All other constraints and rules in LRC are as in (Brennan *et al.*, 1987). The LRC algorithm is as follows:

1. **Preprocessing**: $Cb(U_{n-1})$ and $Cf(U_{n-1})$ are available from previous utterance.

2. **Utterance processing**: Parse and extract incrementally from U_n all references to discourse entities. For each pronoun do:

   (a) Search for an antecedent intrasententially in $Cf\text{-}partial(U_n)$, where *Cf-partial* is a list of all processed discourse entities in U_n that occur before the pronoun. The antecedent must meet feature and binding constraints.

   (b) Search for an antecedent intersententially in $Cf(U_{n-1})$. The antecedent must meet feature and binding constraints.

3. **Creation of *Cf* from current utterance**: Create a Cf-list of U_n by ranking discourse entities of U_n according to grammatical function. Grammatical function is approximated by a left-to-right breadth-first walk of the parse tree.

In addition to introducing incrementality to Centering Theory[3], Tetreault tested the impact of two psycholinguistic claims about Cf ranking on the performance of LRC.

[3] Strube (1998)'s S-list approach, a different formalization of Centering Theory, is another example of an incremental algorithm. Besides incrementality, Strube uses a different ranking function: instead of grammatical roles, discourse entities are ranked on the basis of their information status (Prince, 1981).

The first one concerns ranking of prepended phrases (non-subject surface-initial positions); the other, ranking of the possessor and the possessed entities within complex NPs (following the proposal of Walker and Prince (1996)). Tetreault's results seem to show that prepended phrases should not be ranked prominently, contrary to the suggestion of Gordon *et al.* (1993), while Walker and Prince's hypothesis about linear ranking of complex NPs (from left to right, leftmost as more prominent) performed marginally better than the hypothesis to rank the possessed entity higher than the possessor entity.

Tetreault tested his algorithm on *New York Times* articles (the first two sections of the *Wall Street Journal* corpus) and a corpus of fictional texts. The best instantiation of LRC achieved a success rate of 80.4% on the WSJ and 81.1% on fictional texts.

3.1.2 An empirically-based system for processing definite descriptions (Vieira and Poesio, 2000)

Vieira and Poesio's system resolves definite descriptions, i.e., definite NPs with the article "the". Other types of definite NPs such as pronouns, demonstrative NPs, or possessive descriptions are not addressed in their approach. (In the text that follows I will be using the terms "definite NPs" and "definite descriptions" interchangeably to refer to definite descriptions.) The system's design was based on their corpus study of definite descriptions in the *Wall Street Journal* corpus (Poesio and Vieira, 1998). That study highlighted the prevalence of discourse-new descriptions, i.e., definite NPs that are not anaphoric (see also (Fraurud, 1992) for a similar finding in Swedish), and therefore the resolution system was designed to handle both discourse-new descriptions and anaphoric definite NPs.

Discourse-new definite descriptions are recognized through a set of heuristics based on research by Hawkins (1978), who identified a number of correlations between certain types of syntactic structure and discourse-new definite descriptions. For instance, definite NPs with *special predicates* such as pre-modifiers "first" and "best" and full relative clauses, e.g., "the first person to sail to America", and heads that take factive complements, e.g., "the fact" in "the fact that there is life on Earth" are mostly discourse-new. Definite NPs in appositive constructions, e.g., "the president of Phillips Petroleum Co." in "Glenn Cox, the president of Phillips Petroleum Co.", and subjects of certain copular constructions, e.g., “the man most likely to gain custody of all this” in “the man most likely to gain custody of all this is a career politician named David Dinkins”, are also discourse-new. Finally, some proper names such as “the United States”, definite NPs with proper noun modifiers such as “the Iran-Iraq war” and references to time, e.g., “the morning”, are also discourse-new and they do not require an antecedent.

Anaphoric definite descriptions fall into two categories: *directly anaphoric* and *bridging descriptions*. Direct anaphora are definite NPs that refer to the same entity as their antecedents and that have same head nouns as their antecedents. Bridging descriptions are definite NPs that either (a) have an antecedent denoting the same entity but using a different head noun, e.g., “the book ... the novel”; or (b) are related by a relation other than identity to an entity already in the discourse model, e.g., “the room ... the chandelier”.[4]

Direct anaphors are resolved by string matching of the head nouns of the anaphor and antecedent NPs. Also, some noun pre- and post-modification is taken into account. For instance, “a blue car” cannot serve as antecedent for “the red car” and “the house on the left” cannot serve as antecedent for “the house on the right”. Taking into account the semantic contribution of the pre- and post-modifiers requires commonsense reasoning, so instead Vieira and Poesio use heuristics: (1) an antecedent and anaphor match if the pre-modifiers of a definite description are a subset of the pre-modifiers of the antecedent; (2) non-premodified NPs can serve as antecedents for any same-head definite; and (3) if both NPs have post-modifiers and the modifiers are different, then the anaphor and the antecedent do not match.

Bridging descriptions require lexical knowledge for their resolution, e.g., that a novel is a book (in the sense of a physical object). To acquire this knowledge, Vieira and Poesio used the WordNet lexical hierarchy, named entity heuristics, and heuristics for compound nouns to resolve examples such as “the stock market crash ... the

[4]Note that Vieira and Poesio’s notion of bridging is wider than the notion I’m using. (And it is quite different from what most of researchers are using.) In Vieira and Poesio’s classification, all other-anaphors would be considered as bridging phenomena. I reserve the term “bridging” for examples such as 6 in Chapter 2 in which the anchor of the other-anaphor is not explicitly mentioned in the text. Instead, the text contains a related entity, and the hearer is invited to construct the anchor from the contextual information and the commonsense knowledge that a receiver is a part of a telephone set.

market".

Both anaphora resolution modules further use discourse segmentation and recency heuristics, to constrain the life span of discourse entities and limit the search space. For direct anaphora, the system searches for antecedents in a four-sentence window. For bridging, the window is set to five sentences. (Both window sizes were determined empirically.) Most recent entities are considered first.

The order in which different processing modules are applied (e.g., should discourse-new descriptions be identified before, after, or in parallel with resolving direct and bridging anaphora?) was determined empirically. Vieira and Poesio tested a hand-crafted decision tree and a decision tree learned by an ID3 classifier (Quinlan, 1993). In general, both decision trees attempt to simultaneously (1) classify a definite NP as discourse-new, directly anaphoric or a bridging description and (2) if it is anaphoric, find its antecedent. The hand-crafted decision tree performs the tests in the following order (the algorithm proceeds to the next step if the current test has failed):

1. Does the definite NP contain a special predicate such as e.g. "fact"? If yes, then it is discourse-new and there is no need to look for an antecedent.

2. Does the definite NP occur in an appositive construction? If yes, it is discourse-new.

3. Else search for an antecedent among all NPs available in the window, by matching head nouns and checking pre- and post-modifiers. If an antecedent is found, the definite description is classified as direct anaphora.

4. Is the head noun of the definite NP a proper noun? If yes, it is discourse-new.

5. Does the definite NP have a restrictive postmodifier? If yes, it is discourse-new.

6. Does the definite NP have a possessive pre-modifier? If yes, it is discourse-new.

7. Does the definite NP occur in a copular construction? If yes, it is discourse-new.

If the tests above failed, the definite NP is likely to be a bridging description, and the system applies proper name heuristics, heuristics for compound nouns, and WordNet lookup (in that order), attempting to find its antecedent.

System's version	P	R	*F*
V1 overall	76%	53%	62%
V2 overall	70%	57%	62%

Table 3.2: Overall performance of Vieira and Poesio (2000)'s system versions 1 and 2 on the test data.

System's tasks	P	R	*F*
Discourse-new identification	72%	69%	70%
Anaphora classification	90%	67%	77%
Anaphora resolution	83%	62%	71%
Overall	76%	53%	63%

Table 3.3: Task-based performance of Vieira and Poesio (2000)'s system version 1 on the test data.

The automatically derived decision tree uses five binary features: special predicate, direct anaphora, apposition, proper noun, and restrictive post-modification. It first checks whether the antecedent has the same head noun as the definite NP. Next, it checks for presence of restrictive post-modifiers, appositive constructions, factive and other special predicates and whether the definite NP is a proper name. The hand-crafted decision tree performed slightly better than the automatically learned decision tree.

The system's performance was evaluated on a development corpus of 1,000 definite NPs and an independent test corpus of 400 definite NPs; both corpora were extracted from the *Wall Street Journal* corpus (Penn Treebank I). Vieira and Poesio presented results for two systems (the data below are for the hand-crafted decision tree): version 1 which does not handle bridging descriptions and version 2 which does (Table 3.2). Furthermore, separate results are available for the task of NP classification (discourse-new, directly anaphoric or bridging) and for the task of anaphor resolution for version 1 of their system (Table 3.3).

While the data in Table 3.2 suggest that version 2 of the system had higher recall but lower precision, there is an important difference in how the two versions of the system were evaluated on the test data: version 2 was evaluated only as a classifier, and the antecedents found by the system were not analyzed. Version 1, on the other hand, was evaluated both as a classifier and as a resolver, with manual inspection of the output.

There is no data on system's performance on the test set for bridging descriptions. There is, however, some information about how well the system performed on the training data (Table 3.4 and Table 3.5). Table 3.4 shows the performance of the system as a classifier. Of the 204 bridging samples in the development corpus it correctly identified 61 NPs as bridging, or 30%. It misclassified as bridging descriptions 100 NPs in the training corpus (so the number of false positives is almost twice the number of true positives).

Vieira and Poesio note that bridging descriptions are the most difficult class to resolve because of the amount of commonsense and world knowledge required for their processing — the system needs both to find the antecedent and to identify a relation that links the antecedent and the anaphor. This dependence on commonsense knowledge means that in general a system can resolve bridging descriptions only when supplied with an adequate knowledge base. Table 3.5 gives an indication of the adequacy of WordNet as a knowledge base for this task on their data. Of the 204 examples of bridging descriptions in the training corpus, the system found WordNet relations for 106 of them. (Note that not all bridging NPs require WordNet for their processing, e.g., proper name antecedents were resolved through heuristics. Also, Vieira and Poesio performed a WordNet search on the whole five-sentence window.) But only 30 antecedents of the 106 for which the system found a WordNet relation were the correct ones (28%). The main reason for this, wrote Vieira and Poesio, is that even if a semantic relation exists in WordNet between an anaphor and antecedent it is not a sufficient condition that the antecedent is actually the correct antecedent for this particular anaphor. In many cases, the text contains a distractor which stands in one of the semantic relations with the anaphor that they consider (synonymy, hyponymy, meronymy, and sister). They drew the conclusion that some sort of focusing must play a crucial role in restricting what entities are possible as antecedents of bridging anaphors. I will return to this issue in Section 3.2 when I introduce the reason for the two algorithms for other-anaphors.

Two other reasons for the poor performance of their WordNet heuristics on bridg-

Bridging class	Found	False positives
Names	12	14
Common nouns	15	10
WordNet relations	34	76
Total	61	100

Table 3.4: Results of manual evaluation of Vieira and Poesio (2000)'s bridging heuristics on the training data.

Bridging class	Relations found	Correct antecedents	% Correct
Synonymy	11	4	36%
Hyponymy	59	18	30%
Meronymy	6	2	33%
Sister	30	6	20%
Total	106	30	28%

Table 3.5: Evaluation of the search for antecedents of bridging descriptions in WordNet (Vieira and Poesio, 2000).

ing descriptions were word sense ambiguity, which was responsible for some false positives, and incompleteness of the lexical information encoded in WordNet. Many synonymy, hyponymy and meronymy relations that Vieira and Poesio encountered in their corpus were not recorded in WordNet: only 46% of corpus relations were observed in WordNet. And even if the relation was implicitly stated in WordNet, it was not always straightforward to retrieve. For instance, consider the following example "the house ... the floor". The concepts "floor", "wall", and "room" are encoded in WordNet as part of "building" but not part of "house", which is a hyponym (more specific concept) of "building". (So, the information is available for the more general term, but not its hyponyms.)

WordNet and sense ambiguity were also responsible for some errors in resolving bridging descriptions with proper name antecedents. Finally, compound noun antecedents such as "a 15-acre plot ... the 15 acres" are hard. (Actually, the antecedent in this example is not a compound noun, but the modifier "15-acre".)

3.1.3 The COCKTAIL system (Harabagiu and Maiorano, 1999)

Harabagiu and Maiorano (1999)'s method for coreference resolution (aptly called COCKTAIL) is a set of heuristics informed by their own and other's research on coreference. The system resolves both pronouns and nominal anaphors, including proper names. Each type of anaphor is resolved by a set of heuristics unique for that anaphor type. For instance, there are separate heuristics for reflexive, possessive, relative, 3rd person and 1st person pronouns; definite, bare and indefinite nominals; and proper names. The heuristics operate on various syntactic, semantic, and discourse cues. For some anaphor types the system further performs semantic checks which combine sortal constraints from WordNet with co-occurrence information from a treebank and conceptual glosses in WordNet. Unlike other systems, the antecedents are sought not only in the preceding text but also in the coreference chains that have already been built by the system. Finally, some heuristics make use of derivational morphology.

To resolve nominal coreference, Harabagiu and Maiorano use the following nine heuristics (here rendered in a somewhat simplified form). The search is performed from right-to-left, first on coreference chains and then on the preceding text. The heuristics are applied in the order in which they are presented.

1. If the anaphor is head of an appositive, resolve the antecedent to the preceding NP;

2. If the anaphor belongs to an NP, search for an antecedent such that it has the same head noun and identical, coreferring, or more specific modifiers.

3. If the anaphor is head of an NP, search for an antecedent which is a proper name with the same head.

4. Search for an antecedent such that it is a proper name with the same NE category as the anaphor, e.g., person, organization or location.

5. Search for an antecedent such that it is a synonym or hyponym (more specific term) than the anaphor.

6. Search for an antecedent which is a definite NP or a modified NP and is semantically consistent with the anaphor.

7. If the head of the anaphor NP or one of its hypernyms or holonyms is a nominalization, then search for a verb *V* deriving this noun or one of its synonyms. Then take as antecedent *V*'s object. This heuristic allows resolution of samples such as Example 46:

   (46) IBM and Mr. York wouldn't discuss his *compensation package* which could easily reach into seven figures. **The subject** is sensitive at a time when IBM is laying off thousands of employees.

   In Example 46, the verb "discuss" has a nominalization "discussion" which is a "communication", which in turn is a hypernym of "subject". The antecedent is then the object of "discuss".

8. If the anaphor is head of a PP preceded by a nominalization, search for a verb *V* which derives this nominalization or one of its synonyms. Resolve the antecedent to the object NP of this verb if the object NP and anaphor NP have the same category (presumably WordNet hypernym).

9. If everything fails, coerce the anaphor to an antecedent which is its hypernym or meronym. (This heuristics attempts to resolve metonymies; it has not been used in evaluation, as the test corpus contained very few metonymic anaphors.)

These heuristics were tested on an unspecified corpus and were reported to have reached a precision of 63%–98%, the lowest precision score for heuristic 8 and the highest for heuristic 1. The paper contains no information about the system's recall.

3.2 Two algorithms for other-anaphora

The remainder of this chapter presents two symbolic algorithms I developed for other-anaphora, LEX and SAL. The algorithms use different types of knowledge: lexical semantics (LEX) and grammatical salience (SAL). (Both algorithms also employ recency constraints.) Using these separate knowledge sources allowed me to compare the two types of knowledge and determine the extent to which they contribute to resolving other-anaphors. While it is known that certain types of grammatical knowledge, e.g., grammatical role of antecedent (and anaphor), play an important role in the resolution of pronominal references (for other factors relevant to pronoun resolution see chapter 18 in (Jurafsky and Martin, 2000)), it is not clear whether grammatical knowledge is important for the resolution of non-pronominal anaphora as well. Vieira and Poesio (2000) hypothesized that some sort of focusing mechanism might be necessary to constrain the set of entities which can serve as antecedents of bridging descriptions (Section 3.1.2). Following this idea, Poesio (2003) investigated the correlation between bridging descriptions and focusing (as formalized within Centering Theory), but found no support for the hypothesis. He reported that choosing as antecedent a backward-looking center of the previous sentence, $Cb(U_{n-1})$ (Section 3.1.1), or a preferred center, $Cp(U_{n-1})$, lead to correct results in 33.6% and 38.2% of all bridging cases respectively. This should be compared with a 44.5% accuracy when resolving the antecedent to a first-mentioned (leftmost) entity in a preceding sentence. For other-anaphora, Bierner (2000) suggested that other-anaphors might be resolved through standard discourse anaphora techniques based on salience (combined with consistency checks). Bierner did not present any evaluation of his hypothesis.[5] Contrary to the two hypotheses above, Strube (2002) claimed that syntactic factors such as grammatical role of antecedent and syntactic parallelism (i.e., when anaphor and antecedent have the same grammatical role), are not important in resolving references of definite NPs.

With respect to semantics, it has been noted by, e.g., Strube and Hahn Strube and Hahn (1999) that nominal anaphora is far more constrained by conceptual criteria than is pronominal anaphora. And even in pronoun resolution, researchers have stressed the importance of semantic information. Tetreault (2001), for instance, wrote that resolving pronoun references on the basis of syntactic information only is naive, and that pronouns would ideally be resolved by a combination of syntax and semantics. Considering that non-pronominal NPs carry much more information than pronouns (in particular, lexical information), this statement intuitively makes sense. This however should be contrasted with the findings of Vieira and Poesio (2000), who came to the conclusion that semantic knowledge (as encoded in WordNet) is not sufficient to resolve their corpus of bridging descriptions. They found that the texts contain many

[5] Bierner's primary focus was on "other" with structurally available antecedents.

distractors, e.g., entities that stand in the same kind of semantic relations with the anaphors as do the correct antecedents. In their corpus, in about half of the cases, a competing discourse entity was "semantically closer" to the bridging anaphor than the correct antecedent. Also, in almost 40% of cases, no semantic relation was found between bridging descriptions and their antecedents. (They only considered the following relations: synonymy, hyponymy, meronymy, and sisterhood.)

So, exactly how much does lexical semantics contribute to the resolution of other-anaphora? And what is the contribution of syntactic constraints such as grammatical salience? These two questions are answered in the remainder of this chapter, by comparing the performance of the LEX and SAL algorithms on a common corpus of examples from the *Wall Street Journal* corpus. The remainder of the chapter is as follows. In Section 3.3 I give the details of the corpus. Section 3.4 presents the details of LEX and Section 3.5 the details of the SAL algorithm. The algorithms are evaluated and compared on a set of examples from the *Wall Street Journal* corpus in Section 3.6.

3.3 Corpus collection and preparation

3.3.1 Data collection

I collected 189 other-anaphors with non-pronominal heads and non-structural antecedents from the *Wall Street Journal* corpus (Penn Treebank release 2, sections 00-02). Each sample contained at least one other-anaphor, in the context of the sentence in which it was used, plus one previous sentence. The samples were extracted using Tgrep, a precursor of Tgrep2, a search engine for parse trees.[6] The samples were extracted in the following fashion. I first extracted all samples with the modifiers "other" and "another", without paying attention to the type of construction in which they occurred. (So, the search returned other-anaphors as well as list-other constructions, idiomatic expressions, discourse connectives and pronominal-like phrases "the other one", etc.) I then wrote a filter which filtered out idiomatic expressions (e.g., "the other week"), reciprocal "each other" and "one another", discourse connectives (e.g., "on the other hand" and "in other words"), elliptic constructions "one X . . . the other(s)" and "one X

[6] http://tedlab.mit.edu/~dr/Tgrep2/

... another", one-constructions "the other one" and "another one", than-comparatives "Xs other than Ys", and list other-anaphors. As a result of this, I had a corpus of 189 examples of other-anaphors with non-structural antecedents.

I used examples from the WSJ, rather than the BNC for three reasons. First, available corpora contain various amounts of "other" and "another". The GNOME corpus, for instance, which has been used to develop general algorithms for generation of nominal expressions[7] and by Poesio and Modjeska (2002) in their study of demonstrative NPs, contains very few occurrences of "other".[8] It is possible that the frequencies of various types of anaphors correlate with the genre of the texts and the communicative goals of their authors. Second, parts of the WSJ corpus have been used as a workbench for training and testing of various pronoun and definite NP resolution system, e.g., (Ge *et al.*, 1998; Tetreault, 2001; Vieira and Poesio, 2000). While other-anaphora is sufficiently different from pronominal and definite NP anaphora to call for a different resolution method (e.g., NPs with "other" and "another" are overwhelmingly anaphoric, while more than half of definite descriptions are first-mention (Fraurud, 1992) and therefore they do not require an antecedent), it is a phenomenon that shares some similarities with definite descriptions, in particular bridging. Finally, the Penn Treebank is parsed, which facilitated corpus preprocessing, and it also allowed efficient and correct computation of the grammatical relations needed for the SAL algorithm. (It is possible to parse unparsed corpora, e.g., BNC, using stand-alone parsing software, but, since no software is perfect, that could have introduced errors which would propagate through the system and affect its performance.)

I used the same corpus of other-anaphors for development of the algorithms and for testing. This might have resulted in some overfitting. However, there exists no corpus with other-anaphors already annotated (the BNC corpus in Chapter 2 did not include "another"; see also the remark above regarding grammatical role extraction).

All NP antecedents of other-anaphors in the corpus were annotated to create a gold standard. As in the case with the BNC study, the annotation was not validated

[7] `http://www.hcrc.ed.ac.uk/~gnome/index\_main.html`

[8] The museum texts in the GNOME corpus describe museum objects and the artists that produced them. The pharmaceutical subcorpus consists of leaflets providing patients with mandatory information about their medicine.

SYNTACTIC POSITIONS THAT **CAN NOT** REALIZE BOTH OTHER-ANAPHORS AND THEIR ANTECEDENTS:

Apposition: (a) NP preceding an appositive, if appositive contains “(an)other”; (b) appositive NP following an other-anaphor:

(47) a. Mary Elizabeth Ariail, **another social-studies teacher**

b. **The other social studies teacher**, Mary Ariail . . .
(both not “other than Mary Elizabeth Ariail”)

Copular clauses: (a) subject NP of a copular clause, if the anaphor is predicate; (b) predicate NP if the other-anaphor is the subject:

(48) a. The reputed wealth of the Unification Church is **another matter of contention**.

b. **The other matter of contention** is the reputed wealth of the Unification Church.
(both not “other than the wealth”)

Possessives S/OF: (a) the possessor NP, if other-anaphor realizes the possessed entity ; (b) possessive PP complement of an other-anaphor:

(49) a. Koito’s **other shareholders**

b. **other shareholders** of Koito
(both not “other than Koito”)

Constructions with spatio-temporal “there”, in which the anaphor is the head of the sentence: e.g.,

(50) a. In London, there are **other locations** where we could meet. (not “other than London”)

b. On Tuesday, there are **other times** when we could meet. (not “other than Tuesday”)

Figure 3.1: Syntactic constraints on antecedents of other-anaphors.

by other annotators. I used my linguistic intuitions, common-sense knowledge, and previous experience and training as a linguist to interprete the anaphors and identify their antecedents. Where my interpretation of an example is debatable, I briefly note that and offer an alternative interpretation.

The corpus contained examples with split antecedents. In such cases, all antecedents in the window of current and previous sentences were annotated.

3.3.2 Data preparation

Not all NPs can serve as antecedents for other-anaphors. There are syntactic environments which cannot realize antecedents of other-anaphora, subject to certain conditions. So far, I have identified four such environments (Figure 3.1). NPs that occurred in these positions were manually removed from the dataset.

NPs containing a possessive modifier, e.g., "Spain's economy" were split into a possessor phrase, "Spain", and a possessed phrase, "economy". Coordinated NPs, e.g., "risk, technology and innovation", were split into their constituent parts using simple heuristics. Next, all sentences in the corpus were processed to extract head nouns of all top-level and embedded NPs. This was done in three steps. First, I extracted base NPs, i.e., NPs that contain no further NPs within them. So, for example, from

```
(S (NP-SBJ (NP (DT the)
               (NN question))
           (PP (IN of)
               (NP (NP (NP (NNS investors)
                                (POS ’))
                           (NN access))
                   (PP (TO to)
                       (NP (DT the)
                           (NNP U.S.)
                           (CC and)
                           (JJ Japanese)
                           (NNS markets))))))
```

I got

```
(NP (DT the) (NN question))
(NP (NNS investors) (POS '))
(NP (NN access))
(NP (DT the) (NNP U.S.) (CC and) (JJ Japanese) (NNS markets))
```

Next, I filtered out empty NPs, e.g., (NP-SBJ (-NONE- *)). And finally, I extracted the head nouns:

```
question
investors
access
markets
```

To avoid deciding which nouns were used as modifiers and which were part of a compound noun, all strings with proper and common noun tags to the right of eventual determiners, quantifiers, and adjectival modifiers were treated as compound nouns.

These procedures resulted in a set of lists which contained head nouns of those NPs that can realize antecedents of other-anaphors. For each anaphor, there were two lists: one with NP heads from the sentence containing the anaphor; the other containing NP heads from the previous sentence.

Before testing the LEX and SAL algorithms, further preprocessing was necessary. When testing the LEX algorithm, I removed from the dataset pronominal NPs, as (1) they do not carry enough lexical information, and (2) they cannot be looked up in WordNet. When testing the SAL algorithm, pronominal NPs were left intact in the corpus. Also, for the LEX experiment only, the order of NPs in each sentence was randomized and named entities were classified according to the scheme in Table 3.6. This scheme was modeled on the MUC-7 Named Entity Task Definition (Chinchor, 1997), with a few differences.[9] Some of the differences are as follows. Unlike MUC-7, I allowed for nested expressions, e.g., "U.S.A." (tagged LOC) in "Campbell U.S.A." below, needed to resolve examples such as the following:[10]

[9] NE annotation was performed manually to avoid error propagation.

[10] This however was used sparingly, so for instance while it is possible to annotate "100 million" in "$ 100 million" as NUM, and the whole phrase as MONEY, I only used the tag MONEY.

(51) The way that we've been managing *Campbell U.S.A.* can hopefully spread to **other areas of the company**. (i.e., "other areas than the U.S.A. branch of Campbell")[11]

Also, the MUC-7 scheme for TIMEX expressions is quite elaborate and includes NPs which are not proper names, e.g., seasons ("autumn") and relative temporal expressions ("last night", "today"). I did not annotate such expressions as they can be resolved by lexical means. Also, I did not use separate tags for date and time of day; both types of entities were tagged TIME. With respect to numerical entities, I annotated all kinds of numerical expressions, while the MUC-7 schema annotates only currency. Finally, names that do not fall into MUC-7 ENAMEX, NUMEX and TIMEX categories, e.g., titles, roles, and non-organizational entities (Table 3.6) were annotated with the tag MISC.

3.4 LEX: a lexical resolution algorithm for other-anaphora

3.4.1 Types of lexical relations that LEX can handle

In Chapter 2, I identified four major types of relations between other-anaphors and their antecedents: `instance-of`, `subset-of`, `same-predicate`, and `redescription`.[12] These relations can be operationalized as hypernymy, same-predicate, metonymy, metaphor, bridging, and redescription. LEX handles only same-predicate and hypernymy relations, both with common and proper name antecedents, e.g., Examples 52 and 53 below.

(52) Employers can pay the subminimum for *90 days*, without restriction, to workers with less than six months of job experience, and for **another 90 days** if the company uses a government-certified training program for the young workers.

(53) Mr. Stoll draws his title from the *cuckoo's* habit of laying eggs in the nests of **other birds.**

[11] It is debatable whether the antecedent in this example is "Campbell U.S.A." or "U.S.A."

[12] Same-predicate is not really a relationship, but rather a convenient conceptual and computational tool. If the same head noun is used in the description of the anaphor as is in the description of the antecedent, I say that the relationship is that of same-predication.

Tag	Explanation	Examples
TIME	Capitalized temporal expressions	“April 30, 1956”, “Wednesday”
PERSON	Named person or family	“Wilbur Ross Jr.”
ORG	Named corporate, governmental, or other organizational entity	“IBM”
PRODUCT	Name of commercial product	“Thunderbird”, “Leche Fresca”
MONEY	Monetary expression	“$ 101 million”
NUM	Numerical expression; neither currency nor time	“45”, “3 1/5”
LOC	Location: name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)	“U.S.”, “Aslacton”
MISC	Other NEs	“Great Depression”, “the President”

Table 3.6: Named entity classes for LEX

LEX focuses on hypernymy and same-predicate, because they are among the most frequent relations. For instance, in the first three sections of the *Wall Street Journal* corpus, on which LEX was tested, common and proper name antecedents related to the anaphor through a same-predicate relation account for 15% of all cases (Table 3.7). As "same-predicate with proper names" I consider the following examples with named entity antecedents: "the Nagymaros dam ... another dam" and "60% of the executives ... the other 40%" ("60%" is tagged as a numerical expression). Hypernymy is by far the most common relation (40%), especially with proper name antecedents (31%). Another large class is redescription, which accounts for 18% of samples, and "inference", which involves mostly examples with implicitly evoked antecedents (10%). "Other relations" (9%) are undecidable samples such as those in which the anaphor occurs in the first sentence of the article, without any obvious antecedent, and phrases that seem to function like genre-specific discourse connectives, e.g., "in other commodity markets". Bridging and metonymy are less common, with 5% and 3% of occurrences, respective.

Another reason for focusing on hypernymy and same-predicate relations was that they seemed easier and cheaper computationally than other types of relations, as they can be extracted automatically, and they did not seem to require reasoning beyond conventional lexical knowledge available from, e.g., WordNet.

3.4.2 LEX: overview of the system

LEX consists of three modules: LEX1, LEX2, and NEM (Named Entity Module). All modules were implemented in Perl. LEX takes as input a corpus prepared as described in Section 3.3.2 and outputs a list of antecedents for each other-anaphor. The modules perform the following tasks.

LEX1 is the module which handles the same-predicate relation. It matches the lemmatised head noun of the anaphor with the lemmatised head noun of all other NPs in the sentence and reports a match if the anaphor and antecedent are evoked by the same predicate, e.g., "90 days ... another 90 days", "invalidity benefits ... other benefits".

LEX2 extracts from WordNet a hyponym tree for the anaphor and matches this tree with the antecedent candidates (actually, their WordNet synsets). It reports a match

Type of relation	No. of occurrences	%
Same predicate with common nouns	22	11%
Same predicate with proper names	7	4%
Common noun hypernymy	18	9%
Proper name hypernymy	59	31%
Redescription with common nouns and non-NP antecedents	30	16%
Redescription with proper names	4	2%
Bridging	10	5%
Metonymy	6	3%
Inference	20	10%
Other relation	17	9%
Total	193	100%

Table 3.7: Semantic relations between antecedent and anaphor in WSJ (sections 00–02).

if the noun under consideration is recorded as a direct or indirect hyponym of the anaphor. To extract information from WordNet, I used Jason Rennie's Perl-to-WordNet interface QueryData, version 1.13.[13] Along with hyponyms, LEX2 extracts synset members, to handle examples such as "the ads … any other U.S. auto advertising". For complex terms, since it is not known in advance whether the left-most word(s) of a term is/are a modifier or part of a compound noun, e.g., "lung cancer deaths" vs. "vice president", LEX2 first looks up the whole string "term1 term2 term3". If the string is not found, the script recursively strips off the leftmost term, "term1", and looks up both "term1" and the remaining string "term2 term3". So, for instance, given the phrase "lung cancer deaths", it first looks up "lung cancer deaths", then "lung" and "cancer deaths", and finally "cancer" and "deaths". This strategy is used both for anaphors and antecedents, and it resolves examples such as 54:

(54) In the torrent of replies that followed, *one woman ringer from Solihull* observed that "the average male ringer leaves quite a lot to be desired: badly dressed, decorated with acne and a large beer-belly, frequently unwashed and unbearably

[13] Available from `http://www.ai.mit.edu/~jrennie/WordNet/`

flatulent in peals." **Another woman** wrote from Sheffield to say that in her 60 years of ringing, "I have never known a lady to faint in the belfry."

Note that a "woman ringer" is not a hyponym of the concept "woman" in WordNet; it is an intersection of the concepts "woman" and "ringer".

The Named Entity Module uses a set of simple heuristics to find antecedents which are evoked by named entities:

- If the head noun of the other-anaphor is the word "year", "month", "week", "day" , "time", in singular or plural, propose as antecedent a noun tagged TIME;
- If the head noun is "product", "wares", "merchandise", "goods", "commodity" or "service" (or their plural forms), propose as antecedent a noun tagged PRODUCT.
- If the head noun is "million", "thousand", "dollar", "yen", "pound", propose a as antecedent a noun tagged MONEY;

With names of persons, organizations, and locations, it is impossible to construct a complete list of predicates. Instead, I used the WordNet lexical hierarchy and the following rules:

- If the head noun of the other-anaphor has synset "location" among its hypernyms, suggest as antecedent a noun tagged LOC;
- If any of the anaphor's hypernyms is "person, individual", suggest a noun tagged PERSON;
- If any of the anaphor's hypernyms is "organization", suggest a noun tagged ORG.

If the head noun of an other-anaphor has several hypernyms, e.g., of the five senses of the word "country" three have the hypernym "location" and one "organization", NEM proposes as antecedents all entities consistent with being a location or organization.

Because of how the heuristics were formulated, in particular those concerning names of locations, persons, and organizations, LEX was able to—incidentally—resolve some metonymies, e.g., Example 55.

(55) *Moscow* has settled pre-1917 debts with **other countries** in recent years at less than face value.

This was possible because Moscow is classified in WordNet as "a national capital ISA capital ... ISA region ISA location". This accidental resolution, besides other things, indicated that, with a richer set of NEM heuristics, it might be possible to resolve other types of metonymic relations (and perhaps other types of semantic relations) using the same approach. The cost of developing, tuning, and testing such heuristics, though, is quite high and should be taken into account.

The three resolution modules are applied in a pipeline[14]:

1. LEX1 (same predicate) on current sentence;
2. NEM on current sentence;
3. LEX1 on previous sentence;
4. NEM on previous sentence;
5. LEX2 (hypernymy) on current sentence;
6. LEX2 on previous sentence.

The search terminates when one of the modules found one or more suitable antecedents. This application order of the resolution modules was derived to maximize results on the training data.

There is no particular mechanism for treating cases with split antecedents. If more than one NP qualifies as antecedent, the algorithm returns all of them.

3.5 SAL: A Centering-based algorithm for resolution of other-anaphora

SAL is an extension of Tetreault's Left-Right Centering (LRC), described in Section 3.1.1. My implementation of LRC differs from the original algorithm in two aspects. First, LRC keeps track of all utterances processed so far (and their forward- and

[14]LEX was tested on a two-sentence window: current and previous.

backward-looking centers), and, if the antecedent is not found in the current sentence, it searches all previous *Cf*-lists, one sentence at time, starting with the previous one. I restricted the search space to two sentences, current and previous. Second, I ignored criteria such as gender and number, which seemed less relevant to the resolution of other-anaphora than to the resolution of pronouns. Arguably, one could consider lexical constraints on other-anaphors being similar to agreement constraints on pronominal anaphors, but the goal of comparing LEX and SAL was to examine the extent to which lexical semantics and grammatical role *independently* contribute to the interpretation of the anaphoric references of other-anaphors. To avoid any possible "contamination", the two algorithms were designed to be as complementary as possible.

Grammatical functions of NPs were approximated on the basis of their position in the parsed trees, by a left-to-right breadth-first walk of the trees. Before testing SAL, the lists of antecedents (actually, their head nouns) were sorted by grammatical function, with subjects ranked higher than objects and objects ranked higher than oblique constituents.

With respect to ranking of entities in possessive NPs, I followed Tetreault, who showed that Walker and Prince (1996)'s approach, which ranks entities in linear order (leftmost entity as more salient than the remainder of the NP) performed marginally better than the opposite theory of Gordon *et al.* (1993).

Consider, for instance, the following example:

(56) While … *the question of investors' access to the U.S. and Japanese markets* may get a disproportionate share of the public's attention, a number of **other important economic issues** will be on the table at next week's talks.

The NPs preceding the anaphor in this example were ranked as follows (topmost entity is the most prominent):

```
question
investors
access
markets
share
```

```
public
attention
number
```

SAL resolves the antecedent to the most salient NP, which in the overwhelming majority of cases happens to be the subject NP. It first attempts to resolve the antecedent in the current sentence, i.e., the sentence containing the anaphor. If the beginning of the sentence has been reached and no antecedent has been found, e.g., the anaphor NP is the subject of the current sentence (Example 57), then the antecedent is resolved to the subject of the previous sentence.

(57) ... Mr. Kwan said the Ninja Turtles could make 1989 a record sales year for *Playmates*.

Other Hong Kong manufacturers expect their results to improve only slightly this year from 1988.

3.6 Evaluation of the LEX and SAL algorithms

Of the 189 samples of other-anaphors from the WSJ corpus, 123 samples were used in the evaluation. These were the samples that (1) had explicit antecedents which were (2) realized as NPs (3) in a two sentence window. The remaining 66 cases were as follows. Eleven samples were classified as *undecidable* in the gold standard corpus. (The annotator could not unambiguously identify the antecedents of these expressions.) Since a human annotator could not resolve these examples, one would not expect that a computational method would be able to resolve them either.

Thirty-one samples violated conditions (1) and (2) above. Twenty samples had *implicitly* realized antecedents, e.g., Example 58 below and Example 9 in Section 1.2, and in 11 samples the antecedents were evoked by non-NP constituents, e.g., a VP (Example 59), a sentence, a clause (Example 60), a text segment, or a verbal form (Example 61):

(58) “What’s he doing?” hissed my companion, who was the only other English-speaking member of the convention and whose knuckles were white. “Attention,” yelled our pilot as our basket plunged into the canal. “You bet attention,”

I yelled back, leaping atop the propane tanks , “I’m wearing alligator loafers!” Our pilot simply laughed, fired up the burner and with **another blast of flame** lifted us, oh, a good 12-inches above the water level.

(59) Without admitting or denying wrongdoing, they consented to findings that they had *inaccurately represented the firm’s net capital, maintained inaccurate books and records*, and made **other violations**.

(60) “The fact of the matter is, *I am a marketer*. That’s **another reason** [for the Backer Spielvogel job]”.

(61) Such individuals, many with young children, are in their prime borrowing years — and, having *borrowed* from the bank, they may continue to use it [the bank] for **other services** in later years.

Further, there were 20 samples with antecedents given outside of the two-sentence window. Also, both LEX and SAL are incremental methods, i.e., they attempt to resolve an anaphor as soon it is encountered, therefore they cannot handle cataphoric examples such as the one below.

(62) What is **another name** for the Roman numeral *IX*?

Finally, three samples contained more than one other-anaphors; such examples are currently out of scope for LEX and SAL algorithms; and one example was an idiom that hadn’t previously been observed, “put another way”.

3.6.1 Results for LEX

LEX correctly resolved 53 to 60 instances of other-anaphors out of the 123 cases with explicitly evoked NP antecedents within a two-sentence window (43% to 49% success rate). The first score in the success rate is a strict score; the second is a lenient score. Scoring was performed as follows. All resolutions were compared to gold standard annotation. In the strict scoring, correct antecedents were those antecedents which were identical to actual antecedents in the gold standard corpus. In lenient scoring, I took into account coreference chains, copular constructions, and anaphors with split antecedents. Consider, for instance, the following example:

(63) *This* is a company that has invested in capacity additions more aggressively than **any other company in the industry** [. . .]

LEX resolved the antecedent to “company”. I do not normally consider predicate nominals to be coreferential with their subjects, but this point of view is not unusual; e.g., the MUC-7 Coreference Task suggests annotating “the President of the United States” in “Bill Clinton is the President of The United States” as coreferential with “Bill Clinton”. And since the sentential subject in Example 63 is a lexically uninformative unit— a pronoun—and the predicative NP is the only source of information for a lexical algorithm, I considered this example correctly resolved, under the lenient scoring schema. A similar approach has been advocated in the machine learning paradigm by Ng and Cardie (2002), who suggested considering as antecedent the *most confident* antecedent, rather than the *closest* antecedent. Specifically, with non-pronominal NPs, Ng and Cardie assume that the most confident antecedent is the closest *non-pronominal* antecedent.

Note that “this” in Example 63 refers to an entity that must have been mentioned by name or description in prior sentences. If the algorithm kept track of coreference chains further back than the two-sentence window, all prior descriptions would have been available for resolution. This is one of the areas where the LEX algorithm can be improved.

In Example 64,

(64) *The computer* can process 13.3 million calculations called floating-point operations every second. *The machine* can run software written for **other Mips computers**, the company said.

LEX resolved the antecedent to “the computer”, rather than “the machine”, which is the most recent reference to that entity. Since both interpretations pick the correct referent, examples such as Example 64 were judged as correctly resolved under lenient scoring.

The third case of examples to which lenient scoring has been applied is illustrated by the following:

(65) (The *March* delivery, which has no limits, settled at 14.53 cents, up 0.56 cent a pound.) The *May* contract, which also is without restraints, ended with a gain

of 0.54 cent to 14.26 cents. The *July* delivery rose its daily permissible limit of 0.50 cent a pound to 14.00 cents, while **other contract months** showed near-limit advances.

In the example above, the antecedent of “other contract months” is a set of referents consisting of “March”, “May”, and “July”. LEX correctly identified “May” and “July”, but not “March”, which is evoked outside of the two-sentence window. It is reasonable to assume that, given a larger window, LEX would have been able to find all antecedents in this example.

In general, when evaluating the results of the LEX algorithm on cases with split antecedents, a full point was awarded if the algorithm found all antecedents and half-point if the algorithm did not find all antecedents.

The 63 samples with other-anaphors that LEX resolved incorrectly or not at all can be divided into the following partially overlapping groups: (1) lexical errors, which show limitations of a *lexical* approach to other-anaphora and/or WordNet as a source of lexical information; (2) errors due to LEX, specifically the heuristics of the Named Entity Module and the order of application of the resolution modules; and (3) semantic phenomena not covered by LEX. The latter class includes samples with metonymic antecedents, samples with bridging inferences, and redescriptions. The three groups are roughly of the same size (Table 3.8); they are considered in turn below. Note that the classification is very rough; oftentimes, a particular example belongs to several groups; see, e.g., the discussion about Example 67 below.

Lexical errors range from ones caused by material missing in WordNet to those induced by sense ambiguity and errors that arose from the semantic vagueness of the anaphor itself. Another principled type of lexical error is associated with pronominal antecedents: pronouns do not contain enough information for a lexical algorithm (e.g., Example 63).

What is missing from WordNet are words, e.g., “markdown”, “Walkman”, “risk factor” and word senses, e.g., Example 66 below:

(66) “This technique is applicable to a wide variety of crops,” he said, and added that some modifications may be necessary to accommodate the peculiarities of

Type of error	Nr of occurrences
Word, sense or relation missing from WordNet	10
Sense ambiguity of the anaphor	1
Unresolved pronominal antecedents	3
Semantic vagueness of the anaphor	8
NEM heuristics	14
Application order of resolution modules	4
Metonymy	5
Bridging inferences	10
Redescriptions	7
Total	63

Table 3.8: Error types and frequencies for LEX.

> each type of crop. He said the company is experimenting with the technique on *alfalfa*, and plans to include cotton and corn, among **other crops**.[15]

The noun "crop" has three senses in WordNet: (1) the yield from plants in a single growing season; (2) the stock or handle of a whip; and (3) a pouch in many birds and some lower animals that resembles a stomach for storage and preliminary maceration of food. In Example 66, the noun "crops" is used in the same sense as "plants" or "species", and this sense is recorded for instance in the Merriam-Webster's Collegiate Dictionary[16]: "a plant or animal or plant or animal product that can be grown and harvested extensively for profit or subsistence, e.g., 'an apple crop', 'a crop of wool'."

Also missing from WordNet is taxonomic information, e.g., that a pension is a form of benefit ("pension ... other benefits"); that a thrift is a financial institution (this information is available for the compound noun "thrift institution", but not "thrift", and as a result the NEM module could not resolve "Columbia ... other thrifts"); that a program trader is an investor ("program traders ... other investors"); and that age is a risk

[15] While it is tempting to view "cotton" and "corn" as two other antecedents of "other crops", I believe that the correct interpretation of Example 66 is as follows. "He said the company plans to include crops other than alfalfa, in particular, cotton and corn."

[16] `http://www.m-w.com/dictionary.htm`

factor ("designer's age . . . other risk factors"). These examples are not straightforward even for humans to process. Some of the knowledge involved in the interpretation of these examples is domain-specific, or it can be new to the reader, and therefore it would required accommodation. For instance, the reader might not know in advance that Columbia is a thrift (institution) or that a program trader is a kind of investor. And age is not always considered as a risk factor. By interpreting the anaphors in these examples and finding their antecedents, the reader acquires new knowledge which she can then add to her existing body of knowledge or reject if she does not agree with it. For instance, not all speakers of English agree that statistics is a form of factual abuse ("statistics and other factual abuse").

The following sample is a canonical example of hypernymy:

(67) Mr. Stoll draws his title from *the cuckoo*'s habit of laying eggs in the nests of **other birds.**

However, sense 3 of "bird" is "dame, doll, wench, skirt, chick, bird (informal terms for a (young) woman)", and because (1) this sense has a hypernym "person, individual", and (2) LEX resolves named entities before hypernymy, LEX resolved "other birds" to "birds other than Mr. Stoll". Errors induced by word senses were also reported by Harabagiu and Maiorano (1999) and Vieira and Poesio (2000). To address this problem, Harabagiu and Maiorano required that the most frequent senses of nouns be promoted. I did not impose such a preference on LEX; and as I will show in Chapter 4, using the most frequent sense is not a good solution at all.

Example 68 illustrates errors associated with the semantic vagueness of the anaphor:

(68) While Mr. Dallara and Japanese officials say *the question of investors' access to the U.S. and Japanese markets* may get a disproportionate share of the public's attention, a number of **other important economic issues** will be on the table at next week's talks.

The noun "issue" (as in "economic issues") is a very general concept that could refer to a variety of discourse entities. Other such general concepts are "thing", "alternative", and "factor". The noun "thing" is particularly uninformative; it also occurs in phrases "among other things" which should, perhaps, be treated as idioms:

(69) That designation would, **among other things**, provide more generous credit terms under which the Soviets could purchase grain.

The second category of errors is associated with the limitations of LEX. In particular, the Named Entity heuristics were inadequate. For instance, NEM could not handle examples in which an organization entity is subsequently referred to by predicates such as "shareholder", "steelmaker", "winner", "bidder", "player", and "major creditor". This is because all of these predicates are classified in WordNet as roles associated with people (they have a hypernym "person, individual", but not "organization"). It seems that an additional heuristic could have taken care of such samples. This heuristic could be formulated as following: "If the head noun of an other-anaphor has synset "person, individual" among its hypernyms, suggest as antecedent any NP tagged ORG." To prevent overgeneration, e.g, resolving the anaphor to both a person and an organization when the correct antecedent is a person, it probably would be necessary to impose an ordering on NEM heuristics, such that the antecedent is first resolved to a person entity and, if such entity cannot be found, to an organization entity.

In Example 70 below, the NE heuristics overgenerated:

(70) [Start of the article] *RMS International Inc., Hasbrouk Heights, N.J.*, facing a cash-flow squeeze, said *it* is seeking **other financing sources** and waivers from debenture holders.[17]

Since all modification was eliminated from the dataset before applying LEX, so that the dataset contained only the head "source", rather than "financial source", and because one of the senses of the noun "source" is "location" (by the way, the most frequent sense), the NEM module incorrectly resolved the antecedent to "Hasbrouk Heights, N.J.", rather than to "RMS International Inc.". Note that a "financing source" is something quite different from a "source" ("the place where something begins", according to WordNet). Only a company or individual can serve as a financial source, and therefore, if the concept "financial source" were included in WordNet, it should have been classified as both a person and an organization.[18] This, however, might be beyond

[17] Some readers consider this example cataphoric on "debenture holders". My interpretation is as follows. A company can seek financing sources internally, e.g., by re-thinking its priorities and moving funds from one account to another or by laying off workers, or it can turn to investors.

[18] The concept has at least two meanings, (1) something or somebody who raises or provides capital, and (2) a person in the finance industry who provides information, as in "one financial source said ...".

the scope of WordNet, which was designed as source of domain-independent information, while the concept of "financing source" seems more applicable in the financial domain. (This concept, for instance, is not included in the Merriam-Webster's Collegiate dictionary either.) In general, contrary to what I said in Section 2.7, modifiers supply not only "additional" information, but their semantic contribution to the interpretation of the anaphor might be essential in finding the correct antecedent. The issue of modification in the interpretation of nominal coreference has been raised by, e.g., Harabagiu and Maiorano (1999) and Vieira and Poesio (2000). However, they did not incorporate the semantics of modifiers in anaphor/antecedent interpretation. Rather, they used modifiers as filters, e.g., if both the anaphor and an antecedent candidate are pre-modified, for there to be a match, the pre-modifiers should be either identical or different, or one of them should be more or less specific than the other. A "red car" and a "blue car" cannot corefer, while a "red car" and a "new car" can. In the examples I considered, the modifiers do not describe a property of an object, but rather contribute compositionally to the meaning of an anaphor. This issue requires further attention and research.

Sometimes, the window contains two entities which both qualify as antecedent:

(71) (Integra-A Hotel & Restaurant Co. said its planned rights offering to raise about $9 million was declared effective and the company will begin mailing materials to shareholders at the end of this week.

Under the offer, shareholders will receive one right for each 105 common shares owned. Each right entitles the shareholder to buy $100 face amount of 13.5% bonds due 1993 and warrants to buy 23.5 common shares at 30 cents a share.) The rights, which expire Nov. 21, can be exercised for $100 each.

Integra, which owns and operates hotels, said that *Hallwood Group Inc.* has agreed to exercise any rights that aren't exercised by **other shareholders**. Hallwood, a Cleveland merchant bank, owns about 11% of Integra.

Both "Integra" and "Hallwood Group Inc." are organizations, and thus both qualify as antecedents. However, the correct antecedent is probably "Hallwood Group Inc." It seems that an additional procedure is necessary to handle such examples. This can be a termination condition: stop after finding one antecedent (which thus would lead to incomplete resolutions in cases with split antecedents), or some additional constraint on antecedent realization, e.g., recency, or the fact that both the anaphor and antecedent occur in indirect speech (the complement clause of "said"). A further empirical study is necessary to determine which of these conditions should be used.

Four errors were due to the order in which the LEX modules were applied. Consider, for instance, the following,

(72) PaineWebber Inc., for instance, is forecasting growth in S&P 500 dividends of just under 5% in *1990*, down from an estimated 11% this year. In **other years in which there have been moderate economic slowdowns — the environment the firm expects in 1990 —** the change in dividends ranged from a gain of 4% to a decline of 1% . . .

LEX first attempts to resolve the antecedent to a same-predicate referent, and only if none is found does it apply NEM. For this reason, the antecedent of "other years" was resolved to "this year"in Example 72, while the correct resolution should have been "1990". A different ordering of LEX modules would probably lead to a correct resolution in this case. Note also that the relative clause "in which there have been moderate economic slowdowns — the environment the firm expects in 1990" would have helped to resolve this example correctly. That resolution, however, would possibly require deeper text understanding techniques than those currently available.

The third type of error comprises examples which involve semantic phenomena not covered by the current version of LEX: metonymies[19], e.g., Examples 73 and 79; bridging references, e.g., Examples 74 and 75; and redescriptions (Examples 76 and 77).

(73) First of America said some of the *managers* will take **other jobs** with First of America.

(74) *Bordeaux's first growths from 1985 and 1986* are $60 to $80 each (except for the smallest in terms of production, Chateau Petrus, which costs around $250!). These prices seem rather modest, however, in light of **other French wines from current vintages**.

[19] Though one example of metonymy, Example 55 in Section 3.4.2, was resolved successfully.

Type of relation	In gold standard	Found	Success rate
Same predicate	22	19	86%
Hypernymy, strict scoring	72	31	43%
Hypernymy, lenient scoring	72	38	53%
Total, strict scoring	94	50	53%
Total, lenient scoring	94	57	61%

Table 3.9: Success rate for LEX for same-predicate and hyponymy relations.

(75) Mrs. Gorman took advantage of *low prices after the 1987 crash* to buy stocks and has hunted for **other bargains** since the Oct. 13 plunge.

(76) The dispute between Eastern and *its pilots* is over a "pay parity" clause in *the pilots'* contract. The clause was part of an agreement in which *pilots* accepted a substantial pay cut as long as **no other labor group** got a raise.

(77) Mr. Achenbaum, who had been considering paring down *his firm* or merging *it* with **another small consulting outfit**, soon agreed.

In Example 73, the antecedent is a metonymy `person-for-the-job`. (Alternatively, it can be viewed as bridging: being a manager means having a job.) In Example 74, "first growth" refers to a *property* of some French wines by virtue of 1855 classification (or later revisions to this classification). In Example 75, a bargain is a transaction which *has* a (low) price. In Example 76, pilots are referred to *collectively* as a labor group. And in Example 77, one *learns new information* about the firm.

The results reported so far are for *all* cases with explicitly evoked NP antecedents, regardless of the type of semantic relationship between the anaphor and its antecedent. As I noted in Section 3.4.1, LEX was designed to only resolve same-predicate and hyponymy relations. If this limitation is taken into account, the success rate goes up dramatically, to an 86% success rate for same-predicate antecedents and a 53% success rate for hyponymic antecedents; the success rate for the whole system rises to 61% (Table 3.9).

3.6.2 Results for SAL

SAL successfully resolved 46 occurrences of other-anaphors (37% of all cases with explicitly evoked NP antecedents within a two-sentence window).

Unlike LEX, SAL is not limited by the quality of semantic resources and/or by other semantic constraints, and therefore it resolved some of the samples that LEX did not resolve, e.g., Example 70 mentioned in the previous section. On the other hand, there were many straightforward samples on which SAL failed, e.g.,

(78) An exhibition of American design and architecture opened in September in *Moscow* and it will travel to **eight other Soviet cities**.

The reason for failure was that the antecedent is locative oblique, and thus it was not chosen as antecedent. (No semantic consistency checks were performed.)

3.6.3 Comparing LEX and SAL

Together, LEX and SAL correctly resolved 72 samples of other-anaphors, or 59% of all cases with explicitly given NP antecedents within a two-sentence window. Of the samples that both methods resolved correctly, LEX resolved 72%. Samples that SAL resolved correctly and LEX did not involve metonymy (Example 79), bridging (Example 80), redescriptions (Example 81), WordNet omissions, and pronominal antecedents.

(79) *Georgia-Pacific*, which went down 2 1/2 Tuesday, lost another 1/2 to 50 3/8. **Other paper- and forest-products stocks** closed mixed.

(80) While *this court ruling* was only on Hammersmith, *it* will obviously be very persuasive in **other cases of a similar nature**.

In Example 80, the antecedent is “it”, referring to “this court ruling”, while the entity to be excluded is “the (legal) case in which this court ruling was made”. To correctly interpret the NP “other cases”, one must introduce a new referent, “the legal case in which the court ruling was made”, through a bridging inference from “this court ruling” and then exclude it from the referential scope of “cases”.

In Example 81,

(81) (*Campbell Soup* jumped 3 3/8 to 47 1/8 as the resignation of R. Gordon McGovern as president and chief executive officer sparked a revival of rumors that *the company* could become a takeover target.) Prudential-Bache Securities boosted the stock's short-term investment rating in response to the departure; analyst John McMillin said he believes *the company* will turn to new management "that's more financially oriented."

Other rumored takeover and restructuring candidates to attract buyers included Woolworth ...

the referent of "Campbell Soup", later referred to by "the company", is characterized as a takeover and restructuring candidate, by virtue of serving as the antecedent of the other-anaphor (reiterating the information from the first sentence of the example, but this time via a nominal predication, rather than a copular sentence). LEX resolved the antecedent to "analyst John McMillin", since the head of the anaphor, "candidate", ISA "person, individual".

Samples that LEX resolved correctly and SAL did not involve antecedents evoked by oblique constituents in the same sentence, e.g., locative adverbials (Example 78), possessive modifiers (Example 82), and temporal adverbials (Example 83); and less salient constituents in a preceding sentence, e.g., (Example 84).

(82) Rather, senior administration officials said that the unexpected meeting was scheduled at *Mr. Bush's* request because of *his* preference for conducting diplomacy through highly personal and informal meetings with **other leaders**.

(83) The documents also said that although the 64-year-old Mr. Cray has been working on the project for more than *six years*, the Cray-3 machine is at least **another year** away from a fully operational prototype.

(84) South Korea registered *a trade deficit of $101 million* in October, reflecting the country's economic sluggishness, according to government figures released Wednesday.

Preliminary tallies by the Trade and Industry Ministry showed **another trade deficit** in October, the fifth monthly set back this year, casting a cloud on South Korea's export-oriented economy.

The finding that LEX resolved 72% of the samples that both LEX and SAL resolved correctly and the performance scores for the two algorithms suggested three things. First, the performance of the algorithms overlap to some extent. Second, since SAL resolves the antecedent to an entity with the highest grammatical role ranking (usually the subject) and fails to produce a correct antecedent in almost 2/3 of the cases, the antecedents of other-anaphors are evoked by less salient entities in 2/3 of the samples in the evaluation corpus. (If saliency is taken to correlate with the grammatical role of a discourse referent. There are alternative views on how to determine salience. For instance, within the Functional Centering approach (Strube and Hahn, 1999), salience is taken to correlate with the information status, following the ideas of Prince (1981). In the original work on Centering Theory (Grosz *et al.*, 1986), Grosz et al. suggested that while grammatical role is a major determinant of the ranking of the forward-looking centers — and thus in the salience status of discourse entities — other factors such as word order, clausal subordination and lexical semantics might affect salience as well.) Grammatical role is therefore less predictive of the antecedent than, e.g., lexical semantics. These results correlate with the observation by Strube (2002), who, presenting the results of Strube *et al.* (2002), reported that grammatical role of the antecedent and syntactic parallelism (another factor closely related to grammatical function) did not seem to contribute to the resolution of definite NPs and proper names in their corpus.

3.7 Summary

In this chapter I presented two symbolic resolution algorithms for other-anaphora. They are the first resolution algorithms specifically designed for other-anaphors. LEX resolves antecedents of other-anaphors on the basis of lexical information available from WordNet, pattern matching, recency, and class information for named entities. Currently, LEX focuses on two types of lexical relations between other-anaphors and their antecedents: same-predicate and hyponymy. Samples with bridging and metonymic antecedents and redescriptions require further research. SAL is based on Centering Theory. It is an extension of Tetreault (2001)'s Left-Right-Centering, the state-of-the-art in pronoun resolution. SAL resolves antecedents of other-anaphors on the basis of grammatical salience and recency.

Both algorithms were tested on a common set of 123 two-sentence samples of other-anaphors from the *Wall Street Journal* corpus. The samples contained non-pronominal other-anaphors with non-structural antecedents. LEX's performance on this dataset was quite good, especially considering the relatively unsophisticated approach. LEX successfully resolved 60 anaphors, or 49% of the samples. If we take into account that LEX was designed to handle only two relations, the success rate was 86% on samples with same-predicate antecedents and 53% on samples with hyponymic antecedents. (Many hyponymic antecedents were resolved incorrectly due to limitations of WordNet as source of semantic information.) SAL correctly resolved 46 anaphors, or 37% of the samples.

An analysis of errors made by LEX suggested that one of LEX's weakest points is the module that processes named entities. Specifically, named entity heuristics were inadequate. Moreover, the ordering of named entity heuristics, and the ordering of LEX modules in general, turned out to be important issues.

Another large group of errors was associated with the limitations of a lexical approach to other-anaphora. First, WordNet was insufficient as a resource for lexical information: some words, word senses, and relationships are missing in WordNet. Second, certain lexical units are insufficient sources of lexical information: they either do not contain enough lexical information (pronouns), or they are too general and semantically vague (e.g., the noun "issue"). These issues have implications for the question posed at the beginning of this chapter: How much do lexical semantics constrain the search for antecedents of other-anaphors? My answer, informed by the experiments reported in this chapter, is as follows: substantially, if sufficient lexical and man-power resources are available. (The latter, e.g., to extend and fine-tune the named entity heuristics.)

Since LEX and SAL use different types of information in finding antecedents of other-anaphors — lexical semantics vs. grammatical role — comparing the performance of these algorithms on a common corpus allowed me to examine the extent to which these types of information independently contribute to resolution of other-anaphora. The comparison suggested that grammatical role does not play a significant role in interpreting other-anaphors: SAL's performance was inferior to the performance of the lexical algorithm. Still, SAL resolved several samples that involved metonymic, bridging, and pronominal antecedents, which LEX did not resolve. Grammatical role thus might be one of the several factors to use when resolving such cases in the absence of a dedicated treatment of these phenomena. Other factors to consider for these and other samples are the linguistic form of anaphor and antecedent, gender, number, distance in words and sentences between anaphor and antecedent NPs, syntactic parallelism, etc. The next chapter examines these and other factors in detail, using a machine learning approach.

Chapter 4

A Machine learning approach to other-anaphora

4.1 From LEX and SAL to ML

In the previous chapter I investigated the relative contribution of two types of knowledge in resolution of antecedents of other-anaphora: lexical semantics and grammatical salience. Lexical semantics showed to be more predictive of correct antecedents than grammatical salience: the lexical algorithm LEX outperformed the salience-based algorithm SAL by 32% (12 percentage point difference in success rate). However, other types of information, e.g., the gender and number of anaphor and antecedent, the distance between them, their agreement in grammatical function and semantic class, etc., might play an important role as well. Consider, for instance, again Example 85:

(85) Mr. Stoll draws his title from the *cuckoo's* habit of laying eggs in the nests of **other birds.**

The LEX algorithm failed to correctly resolve this example because all senses of the word "bird" were considered and because one of its senses is the somewhat old-fashioned sense "dame, doll, wench, skirt, chick, bird", informal term for a (young) woman. Clearly, the algorithm would benefit from word sense disambiguation. Alternatively, some researchers in coreference resolution have suggested that only the most frequent senses should be used. Arguably, there is a third way—through imposing a gender constraint on the antecedent, in addition to the lexico-semantic constraints that the algorithm is already employing. Such a constraint would eliminate “Mr. Stoll” from the list of antecedent candidates, as the anaphor is neuter in gender (and so is the correct antecedent).

Other types of information might also prove useful in resolving samples with vague anaphors (“other issues”, “other alternatives”), pronominal antecedents, and bridging and redescription cases. Although lexical information is more useful than grammatical role information, the WSJ corpus in Chapter 3 contained many examples which the LEX algorithm could not resolve for several reasons. First, WordNet does not contain several concepts, word senses and many relations that would be necessary to resolve the samples (Section 3.6.1). For instance, it omits the concept “thrift” (although it contains the concept “thrift institution”), and therefore the anaphor “other thrifts” in Example 86 could not be resolved. This example, however, was correctly resolved by SAL.

(86) *Columbia* won’t comment on all the speculation. But like **other thrifts**, it’s expected to seek regulators’ consent to create a distinct junk-bond entity.

Second, vague anaphors such as “other issues” and “other alternatives” and pronominal antecedents require a different kind of information, as they are lexically uninformative. Again, since SAL is not dependent on semantic information, it successfully resolved several of such cases. Third, samples with bridging and metonymic antecedents and redescription examples (not yet handled by LEX) clearly require a huge amount of domain-specific and domain-independent taxonomic and general world knowledge. Much of this knowledge is not available in WordNet and/or it may be hard to retrieve (Section 3.1.2). As a result, a purely semantic approach to their resolution would probably fail. (Needless to say, a dedicated treatment must be developed for each of these types of antecedents; an enormous research task in itself.) A combination of several types of knowledge (grammatical, semantic, recency, etc.) might (1) eliminate the need for a dedicated treatment and (2) make up for some of WordNet limitations. However, it is not clear how many features would suffice for the task (one would want to keep the processing effort at as minimal a level as possible), or how these various features interact with each other. I attempted one such analysis in Chapters 2 and 3.

However, much of the interaction of the features is not straightforward, because they seem to be preferences rather than absolute constraints such as the syntactic constraints on antecedent realization I presented in Section 3.3.2, and as such they cannot be fully determined by hand or on the basis of one's linguistic intuitions.

Finally, the application order of resolution modules must be determined. Should same-predicate antecedents be resolved first, before testing for hyponymy? Should the system resolve antecedents to NEs before considering common noun antecedents? (See also the discussion in Section 3.6.1 about ordering of NE heuristics.) The more heuristics and features a system uses, the more difficult it becomes to determine the order of their application by hand. This and the other considerations above make it a good case for a machine learning approach. Machine learning approaches require minimal or no supervision; they determine feature relevance on the basis of their distribution in the corpus and the extent to which choosing a particular feature would lead to a correct resolution. Also, they can handle many features and can *learn* how the features interact with each other. (The approach I mostly use, Naive Bayes, does not actually learn how features interact with each other; it assumes one specific interaction, namely it uses all available features and assumes that they are independent of each other.)

In this chapter, I present a machine learning approach to other-anaphora based on the Naive Bayes classifier. The approach was informed by machine learning research in coreference resolution, a related though different task, and I will start with reviewing this work in Section 4.2. I will then introduce the corpus (Section 4.3.1), the features (Section 4.3.2), and the learning framework (Section 4.3.3). The contribution of the features I used, their relevance for resolving other-anaphors and results are discussed in detail in Sections 4.3.4–4.4.2.

4.2 Machine Learning approaches to nominal coreference

4.2.1 (Aone and Bennett, 1995)

Aone and Bennett described a machine learning approach to coreference resolution in Japanese. They trained a C4.5 decision tree learner (Quinlan, 1993) on a corpus of newspaper articles about joint ventures. Their approach used 66 features, of which they explicitly mention only a few: category (possibly POS-tags), grammatical role, semantic class, and distance between anaphor and antecedent. Some of the features they used seem domain- and language specific. They trained several variants of their classifier; the best of them is reported to have achieved an F-measure of 77.4%, having reached a plateau at 250 training documents.

4.2.2 RESOLVE (McCarthy and Lehnert, 1995)

McCarthy and Lehnert's system RESOLVE arose in an attempt to improve the performance of the IE system that UMass used in MUC-3, MUC-4, and MUC-5 competitions. The earlier coreference resolution module of the system used a set of manually engineered rules, which tended to be very conservative, but above all, they were very complex. Rather than modifying and maintaining the existing rules, McCarthy and Lehnert set out to test a decision rule classifier (C4.5). The classifier was trained on the MUC-5 English Joint Venture (EJV) corpus (MUC-5, 1993). They used 8 features, summarized in Table 4.1. Four of the features were NP-level features (NAME and JV-CHILD were used for both the antecedent and the anaphor), and the remaining 4 features were coreference level features. It was subsequently pointed out by e.g., Soon *et al.* (2001), that 3 of the 8 features that McCarthy and Lehnert used were domain-specific, in particular the features JV-CHILD and BOTH-JV-CHILD, though the patterns they used to induce feature values from the corpus seem domain-independent. One such pattern is the construction "company-name-1 OF company-name-2", as in Example 87, which allows one to infer that "company-name-1" is a subsidiary (child) of "company-name-2".

Level	Feature	Description	Values
NP	NAME	Does antecedent contain a name?	yes, no
NP	NAME	Does anaphor contain a name?	yes, no
NP	JV-CHILD	Does antecedent refer to a joint venture child?	yes, no, unknown
NP	JV-CHILD	Does anaphor refer to a joint venture child?	yes, no, unknown
Coref	ALIAS	Does either NP contain an alias of the other?	yes, no
Coref	BOTH-JV-CHILD	Do both NPs refer to a join venture child?	yes, no
Coref	COMMON-NP	Do the anaphor and antecedent share a common NP?	yes, no
Coref	SAME-SENTENCE	Are both NPs in the same sentence?	yes, no

Table 4.1: Features used in the RESOLVE system (McCarthy and Lehnert, 1995).

Level	Feature	Description	Values
NP	DIST	Distance between antecedent and anaphor measured in sentences	0, 1, 2, 3, ...
NP	I_PRONOUN	Is antecedent a pronoun?	true, false
NP	J_PRONOUN	Is anaphor a pronoun?	true, false
NP	DEF_NP	Is anaphor a definite NP?	true, false
NP	DEM_NP	Is anaphor a demonstrative NP?	true, false
Coref	STR_MATCH	Does the string of antecedent matches that of the anaphor having removed determiners?	true, false
Coref	NUMBER	Do antecedent and anaphor agree in number?	true, false
Coref	SEMCLASS	Do antecedent and anaphor belong to the same semantic class, e.g., both are “person”?	true, false, unknown
Coref	GENDER	Do antecedent and anaphor agree in gender?	true, false, unknown
Coref	PROPER_NAME	Are both antecedent and anaphor proper names?	true, false
Coref	ALIAS	Is antecedent an alias for the anaphor, or vice versa?	true, false
Coref	APPOSITIVE	Is anaphor in apposition to the antecedent?	true, false

Table 4.2: Features used by Soon *et al.* (2001).

(87) Familymart Co. of Seibu Saison Group will open a convenience store in Taipei Friday ...

The RESOLVE system was trained and tested using a 50-fold cross validation. It achieved an average F-measure of 86.5% with equal weights given to precision and recall. It should be pointed out, though, that the MUC-5 data set is highly domain-specific: it only involves entities that are companies, governments, or people who entered business joint ventures. With appropriately designed features it is possible to achieve a rather high performance.

4.2.3 (Soon *et al.*, 2001)

Soon et al. presented a coreference resolution system trained on the MUC-6 and MUC-7 data sets (MUC-6, 1995; MUC-7, 1998). The system uses 12 features (Table 4.2), of which 5 are NP-level features and 7 are coreference level features. The features that need additional explanation are SEMCLASS and ALIAS. The feature SEMCLASS has the following values: female, male, person, organization, location, date, time, money, percent, and object. These classes are arranged in a simple ISA hierarchy: male and female classes are children of the class person, while the remaining classes are children of the class object. For each NP, its semantic class was derived from WordNet, specifically, it was the most frequent sense of the head noun of that NP. Semantic classes of the antecedent and anaphor were compared and the value "true" was returned if (1) both NPs were of the same semantic class, e.g., "Mr. Lim" and "he" are both of class male, or (2) one of the NPs was the parent of the other, e.g., in "Mr. Lim ... chairman", "chairman" is a class person, and class person subsumes class male. If the NPs belong to different parts of the network, e.g., "IBM" ISA organization and "Mr. Lim" is male, the value was set to false. If the semantic class of either NP was unknown, then the head nouns of both NPs were compared (by simple string comparison). If they matched, the value was true; otherwise, the value was unknown.

The ALIAS feature was computed for named entities. The alias module used different strategies depending on the named entity type. For dates, e.g., "01-08" and "Jan. 8", it used string comparison. For persons, such as "Mr. Simpson" and "Bart Simpson", the last words of the NPs were matched. For organization names, the alias module performed the last-word comparison and also checked for acronyms, e.g., "International Business Machines Corp." and "IBM".[1]

The training examples were generated as following. In a coreference chain NP1 - NP2 - NP3 - NP4 (available from the annotated corpus), each pair of immediately adjacent NPs, i.e., NP1 and NP2, NP2 and NP3, and NP3 and NP4, created a positive training example. The first NP in such a pair was considered the antecedent; the second the anaphor. Every NP which intervened between the anaphor and its antecedent was paired with the anaphor to create a negative instance. For instance, if NP5 appeared between NP3 and NP4, a negative instance NP5 - NP4 was created.

The system was trained on 30 documents from the MUC-6 corpus and 30 documents from the MUC-7 corpus (a separate classifier for each year). The C5.0 pruning parameters were determined by performing a 10-fold cross validation on the whole training set for each MUC corpus. For MUC-6, the pruning confidence was set to 20% and the minimum number of instances per leaf node to 5. For MUC-7, the pruning confidence was set to 60% and the minimum number of instances to 2.

The system was tested on 30 documents from the MUC-6 corpus and 20 documents from the MUC-7 corpus. For both corpora, the classifier reached a plateau after 25 training documents, with an F-measure of 62.6% for MUC-6 and 60.4% for MUC-7.

Soon et al. also reported how much each of the 12 features contributed to the overall performance of the system. To find out how useful each of the features was, they trained and tested the classifier using one feature at at time. The most informative features for both corpora turned out to be ALIAS, STR_MATCH, and APPOSITIVE; the 9 remaining features together contributed only 2.3% F-measure on the MUC-6 corpus and 1% on the MUC-7 corpus.

While the performance of this system was much lower than that of RESOLVE and the system designed by Aone and Bennett (1995), it is important to point out that Soon et al.'s system is domain-independent. Also, some of the errors arose as a result of errors made by modules executed before the coreference resolution module, e.g., the NP extraction, POS assignment, and named entity classification modules.

[1] It is not clear whether Soon et al. used a gazetteer.

Level	Feature	Description
Lex	SOON_STR	C if, after discarding determiners, the string denoting NP_i matches that of NP_j; else I.
Gram	PRONOUN_1*	Y if NP_i is a pronoun; else N.
Gram	PRONOUN_2*	Y if NP_j is a pronoun; else N.
Gram	DEFINITE_2	Y if NP_j starts with the word "the"; else N.
Gram	DEMONSTRATIVE_2	Y if NP_j starts with a demonstrative such as "this", "that", "these", or "those"; else N.
Gram	NUMBER*	C if the NP pair agree in number; I if they disagree; NA if number information for one or both NPs cannot be determined.
Gram	GENDER*	C if the NP pair agree in gender; I if they disagree; NA if gender information for one or both NPs cannot be determined.
Gram	BOTH_PROPER_NOUNS*	C if both NPs are proper names; NA if exactly one NP is a proper name; else I.
Gram	APPOSITIVE*	C if the NPs are in an appositive relationship; else I.
Sem	WNCLASS*	C if the NPs have the same WordNet semantic class; I if they don't; NA if the semantic class information for one or both NPs cannot be determined.
Sem	ALIAS*	C if one NP is an alias of the other; else I.
Pos	SENTNUM*	Distance between the NPs in terms of the number of sentences.

Table 4.3: Features from Soon *et al.* (2001)'s system used by Ng and Cardie (2002). "Lex" stands for lexical, "Gram" for grammatical, "Sem" for semantic, "Pos" for positional. "C" stands for compatible, "I" for incompatible, and "NA" for not applicable. $*$-ed features were used in the hand-selected feature set for at least one classifier/dataset combination.

4.2.4 (Ng and Cardie, 2002)

Ng and Cardie's system extended the work of Soon *et al.* (2001). Ng and Cardie used a richer set of features (53 in total; Tables 4.3 and 4.4) and made changes in the machine learning framework. They used RIPPER, an inductive rule-learner (Cohen, 1995), and C4.5, a decision tree learner.

The 41 additional features (Table 4.4) were not derived empirically from the corpus, but rather were based on common-sense and the researchers' linguistic intuitions. Among the new features are, e.g., MODIFIER, MAXIMALNP, EMBEDDED, IN_QUOTE, and PARANUM. The MODIFIER feature compares prenominal modifiers of the antecedent and anaphor. It takes values C (for compatible) and I (for incompatible). The anaphor and antecedents are considered compatible if the prenominal modifiers of one NP are a subset of the prenominal modifiers of the other, and incompatible otherwise. The MAXIMALNP feature compares NPs maximal projections; it takes the value I if both NPs have the same maximal NP projection, and C otherwise. The features EMBEDDED_1 and EMBEDDED_2 test for embedding. The features IN_QUOTE_1 and IN_QUOTE_2 test for whether an NP is part of a quoted string. The latter four features are binary "yes"/"no" features. The PARANUM feature measures the distance between the NPs in terms of the number of paragraphs. Soon et al.'s feature STR_MATCH was split in three features, one each for pronominal, proper name, and common noun NPs.

Like Soon et al., Ng and Cardie trained and tested their classifiers on the MUC-6 and MUC-7 coreference corpora. The F-measures they obtained for MUC-6 were 63.8% for C4.5 and 64.5% for RIPPER; for the MUC-7 data, the scores were 61.6% for C4.5 and 61.2% for RIPPER.

A closer examination of the results showed that many of the 53 features performed badly on common nouns. No analysis has been presented of how much individual features contributed, but Ng and Cardie reported that classifiers induced a number of low precision rules for common nouns; e.g., one of the rules covered 38 examples, but had 18 exceptions. Also, they remarked that the feature set they used might have been insufficient for common noun resolution. Another potential source of problems might have been the size of the feature set. Because they added many features without increasing the size of the training set, this might have led to data fragmentation.

Level	Feature	Description
Lex	PRO_STR*	C if both NPs are pronominal and are the same string; else I.
Lex	PN_STR*	C if both NPs are proper name and are the same string; else I.
Lex	WORDS_STR	C if both are non-pronominal and are the same string; else I.
Lex	SOON_STR_NONPRO*	C if both are non-pronominal and the string of NP_i matches that of NP_j; else I.
Lex	WORD_OVERLAP	C if the intersection between the content words in NP_i and NP_j is not empty; else I.
Lex	MODIFIER	C if the prenominal modifiers of one NP are a subset of the prenominal modifiers of the other; else I.
Lex	PN_SUBSTR	C if both NPs are proper names and one NP is proper substring (w.r.t. content words only) of the other; else I.
Lex	WORDS_SUBSTR	C if both NPs are non-pronominal and one NP is a proper substring (w.r.t. content words only) of the other; else I.
Gram	BOTH_DEFINITES	C if both NPs start with “the”; I if neither start with “the”; else NA.
Gram	BOTH_EMBEDDED	C if both NPs are prenominal modifiers; I if neither are prenominal modifiers; else NA.
Gram	BOTH_IN_QUOTES	C if both NPs are part of a quoted string; I if neither are part of a quoted string; else NA.
Gram	BOTH_PRONOUNS*	C if both NPs are pronouns; I if neither are pronouns; else NA.
Gram	BOTH_SUBJECTS	C if both NPs are grammatical subjects; I if neither are subjects; else NA.
Gram	SUBJECT_1*	Y if NP_i is a subject; else N.
Gram	SUBJECT_2*	Y if NP_j is a subject; else N.
Gram	AGREEMENT*	C if the NPs agree in both gender and number; I if they disagree in both gender and number; else NA.
Gram	ANIMACY*	C if the NPs match in animacy; else I.
Gram	MAXIMALNP*	I if both NPs have the same maximal NP projection; else C.
Gram	PREDNOM*	C if both NPs form a predicate nominal construction; else I.
Gram	SPAN*	I if one NP spans the other; else C.
Gram	BINDING*	I if NPs violate conditions B or C of the Binding Theory; else C.
Gram	CONTRAINDICES*	I if the NPs cannot be co-indexed based on simple heuristics; else C. For instance, two non-pronominal NPs separated by a preposition cannot be co-indexed.
Gram	SYNTAX*	I if the NPs have incompatible values for the BINDING, CONTRAINDICES, SPAN or MAXIMALNP constraints; else C.
Gram	INDEFINITE*	I if NP_j is an indefinite and not appositive; else C.
Gram	PRONOUN	I if NP_i is a pronoun and NP_j is not; else C.
Gram	CONSTRAINTS*	C if the NPs agree in GENDER and NUMBER and do not have incompatible values for CONTRAINDICES, SPAN, ANIMACY, PRONOUN, and CONTAINS_PN; I if the NPs have incompatible values for any of the above features; else NA.
Gram	CONTAINS_PN	I if both NPs are not proper names but contain proper names that mismatch on every word; else C.
Gram	DEFINITE_1	Y if NP_i starts with “the”; else N.
Gram	EMBEDDED_1*	Y if NP_i is an embedded noun; else N.
Gram	EMBEDDED_2	Y if NP_j is an embedded noun; else N.
Gram	IN_QUOTE_1	Y if NP_i is part of a quoted string; else N.
Gram	IN_QUOTE_2	Y if NP_j is part of a quoted string; else N.
Gram	PROPER_NOUN	I if both NPs are proper names, but mismatch on every word; else C.
Gram	TITLE*	I if one of both of the NPs is a title; else C.
Sem	CLOSEST_COMP	C if NP_i is the closest NP preceding NP_j that has the same semantic class as NP_j and the two NPs do not violate any of the linguistic constraints (AGREEMENT through SYNTAX); else I.
Sem	SUBCLASS	C if the NPs have different head nouns but have an ancestor-descendent relationship in WordNet; else I.
Sem	WNDIST	Distance between NP_i and NP_j in WordNet (using the first sense only) when they have an ancestor-descendent relationship but have different heads; else infinity.
Sem	WNSENSE	Sense number in WordNet for which there exists an ancestor-descendent relationship between the two NPs when they have different heads; else infinity.
Pos	PARANUM	Distance between the NPs in terms of the number of paragraphs.
Other	PRO_RESOLVE*	C if NP_j is a pronoun and NP_i is its antecedent according to a naive pronoun resolution algorithm; else I.
Other	RULE_RESOLVE	C if the NPs are coreferent according to a rule-based coreference resolution algorithm; else I.

Table 4.4: Features used by Ng and Cardie (2002). “Lex” stands for lexical, “Gram” for grammatical, “Sem” for semantic, “Pos” for positional, and “Other” for other. $*$-ed features were used in the hand-selected feature set for at least one classifier/dataset combination.

In the next step, Ng and Cardie evaluated a version of the system which employed a smaller set of manually selected features (between 22 and 26 features, marked by an asterisk in Tables 4.3 and 4.4). A special effort was made to exclude features which resulted in low precision for common nouns. The slimmer version of the system did lead to an increase in precision for common nouns, but the precision for pronoun resolution on the MUC-7/C4.5 dropped dramatically (from 62.1% to 54.5%), and also the overall recall scores fell 2.8%–8.1%. Nevertheless, the system's performance was still better than that of Soon *et al.* (2001): the F-measures for the MUC-6 data were 69.1%–70.4% (C4.5 vs. RIPPER) and for the MUC-7 data, 63.4%–63.1% (C4.5. vs. RIPPER).

The modifications that Ng and Cardie proposed to the machine learning framework concerned the question of what should be considered as antecedent. Rather than settling on the *first* NP which matches all constraints and preferences, as Soon et al. did, Ng and Cardie proposed searching for *the most likely antecedent*. This approach had implications for training set creation. Instead of generating a positive training instance between an anaphoric NP and its *closest* antecedent, Ng and Cardie generated a positive training instance for its *most confident* antecedent.[2] Specifically, for a non-pronominal NP, the most confident antecedent was assumed to be the closest non-pronominal antecedent (i.e., intermediate pronominal references were skipped). For pronouns, the most confident antecedent was assumed to be its closest preceding antecedent.

These changes in the learning framework, together with the modification of Soon et al.'s STR_MATCH feature, resulted in small but statistically significant gains in performance. (The F-scores above account for these changes.)

4.2.5 (Strube *et al.*, 2002)

Strube et al. reported on experiments in coreference resolution on German data. They used a corpus of 242 short texts (36,924 total tokens) about sights, historic events, and persons in Heidelberg. They used the C5.0 decision tree classifier with standard

[2] Ng and Cardie use the term "most likely antecedent" for describing the coreference clustering algorithm and the term "most confident antecedent" for describing data generation.

Level	Feature	Description	Values
Doc	DOC_ID	Document number	1 . . . 250
NP	ANTE_GRAM_FUNCTION	Grammatical function of antecedent	subject, object, other
NP	ANTE_NP_FORM	Surface form of antecedent	defNP, indefNP, personal, demonstrative, or possessive pronoun, proper name
NP	ANTE_AGREE	Antecedent's person, gender, and number	
NP	ANTE_SEMANTICCLASS	Semantic class of antecedent	human, concrete object, abstract object
NP	ANA_GRAM_FUNC	Grammatical function of anaphor	subject, object, other
NP	ANA_NP_FORM	Surface form of anaphor	defNP, indefNP, personal, demonstrative, or possessive pronoun, proper name
NP	ANA_AGREE	Anaphor's person, gender, and number	
NP	ANA_SEMANTICCLASS	Semantic class of anaphor	human, concrete object, abstract object
Coref	WDIST	Distance between anaphor and antecedent in words	1 . . . n
Coref	DDIST	Distance between anaphor and antecedent in sentences	0, 1, >1
Coref	MDIST	Distance between anaphor and antecedent in markables	1 . . . n
Coref	SYN_PAR	Anaphor and antecedent have same grammatical function	yes, no
Coref	STRING_IDENT	Anaphor and antecedent consist of identical strings	yes, no
Coref	SUBSTRING_MATCH	One string contains the other	yes, no

Table 4.5: Features used by Strube *et al.* (2002).

settings for pre- and post-pruning and Soon *et al.* (2001)'s method for generation of the training corpus.

In designing the feature set, special attention was paid to features which were (1) relevant according to previous research, (2) cheap and robust, i.e., they could be annotated (semi-)automatically, and (3) domain-independent. This resulted in 15 features (Table 4.5), of which one feature was a document-level feature, 8 were NP-level features, and 6 were coreference-level features.

The system was trained and tested using 10-fold cross validation; it achieved an F-measure of 59.97% (equally weighted between precision and recall). Unsatisfied with the results, Strube et al. examined the performance of the features. The most important feature turned out to be the NP form of the anaphor. This led them to a hypothesis that considerable differences in performance can be expected from the classifier with respect to the NP form of the anaphor. To test the hypothesis, the data set was split into subsets, and the classifier was trained independently on each of these subsets. The results confirmed the hypothesis: the classifier performed poorly on definite NPs and demonstrative pronouns (an F-measure of 15.84% and 15.38%, respective), moderately on proper names (an F-measure of 65.14%), and quite good on personal and possessive pronouns (an F-measure of 82.79% and 84.94%, respective). Since definite NPs accounted for more than a third of all positive examples in their corpus (38.19%), Strube et al. felt it was necessary to try to improve their resolution. In particular, the recall was very low, 8.71%, so they focused on raising the recall without losing too much precision. An examination of the data samples suggested that the STRING_IDENT and SUBSTRING_MATCH features were too strong. To balance them, Strube et al. added two new features, ANTE_MED and ANA_MED, minimal edit distance, which measured the similarity of strings by computing the minimal number of editing operations (substitutions, insertions, and deletions) needed to transform one string into the other. The inclusion of the MED features led to a significant improvement in performance: the F-measure for definite NPs rose by 18%, for proper names by 11%; for pronouns the results were unchanged; and the overall F-measure rose by 8% to 67.98%.

4.2.6 Coreference as clustering (Cardie and Wagstaff, 1999)

Cardie and Wagstaff's approach is the only approach known to me which uses *unsupervised* machine learning. It rests on a hypothesis that all NPs used to describe a specific entity or concept are related in some way and that the conceptual "distance" between them is thus small, or smaller than between NPs describing different entities. Therefore, it is possible to view the task of coreference as that of partitioning, or clustering, NPs into equivalence classes. Specifically, given a description of each NP and a method for measuring the distance between a pair of NPs, a clustering algorithm can either put them into the same partition if they are close enough, or in two different partitions if the distance between them is greater than a certain threshold. NPs placed into the same partition are considered coreferent. NPs placed into different partitions are considered not coreferent. The distance is defined as a sum of incompatibility functions, which compare NPs in each given pair with respect to the features in Table 4.6. For example, the function Gender compares the gender of both NPs and returns 1 if they do not match and 0 otherwise. All feature values were computed automatically without any manual tagging. The clustering radius threshold was obtained from their corpus, but, as Cardie and Wagstaff remark, it might be constant across corpora and thus would not need to be recalculated.

The approach was tested on the MUC-6 corpus. It achieved an F-measure of 53.6%, with equal weights to precision and recall. This performance may seem modest, but it important to remember that Cardie and Wagstaff's approach is fully automated and unsupervised. And like Soon *et al.* (2001), they attribute some of the errors to erratic or insufficient output from the preprocessing modules. Specifically, the semantic class module was responsible for many errors: the semantic class distinction was coarse and often inadequate. Also, they wished they had access to a better named entity finder and information about grammatical and thematic roles. Lastly, some of the errors arose from the greedy nature of the clustering algorithm. Cardie and Wagstaff did not report a performance analysis of their algorithm with respect to the features they used.

Level	Feature	Description	Values
NP	Individual Words	Words contained in NP	
NP	Head Noun	The last word in NP	
NP	Position	Unique number from beginning of document	
NP	Pronoun type		nominative, accusative, possessive, ambiguous, none for all other NPs
NP	Appositive	Is the NP an appositive?	yes, no
NP	Number		plural, singular
NP	Proper name	Is the NP a proper name?	yes, no
NP	Semantic class	What is WordNet definition for head noun or its hypernym? A separate algorithm for numeric expressions, money, and companies	time, city, animal, human, object, number, money, company, plus a number of idiosyncratic WordNet concepts e.g., “payment”
NP	Gender	Determined from WordNet and a list of common first names	masculine, feminine, either, neuter
NP	Animacy		animate, inanimate

Table 4.6: Features used by Cardie and Wagstaff (1999).

4.2.7 A competition approach by (Connolly *et al.*, 1997)

Connolly et al. presented a novel approach to coreference resolution, radically different from most of the research in the field.[3] Rather than considering one antecedent candidate at a time, they applied classifiers to successive *pairs* of candidates, each time choosing and retaining the best, until all candidates in the article were considered, thus introducing an element of *competition*. Interested in determining empirical performance baselines for machine learning approaches to the task of reference resolution, they considered four well-known classifiers: a posterior classifier, a Naive Bayesian classifier, a decision tree learner (C4.5), and a Neural Network classifier. They further proposed and tested three hybrid classifiers and one hand-crafted symbolic algorithm. The hand-crafted algorithm consisted of a decision list with approximately 50 entries covering both pronominal anaphors and definite NPs. Hybrid classifiers included a hybrid Bayesian classifier, a hybrid decision tree, and a hybrid Neural Network classifier. All hybrid classifiers attempted to address the issue of high dimensionality by assuming feature independence on data partitions/subspaces. For instance, for the hybrid Bayesian classifier, the features were partitioned into to groups: (1) the agreement features (see below) and (2) all other features. The features in each of the groups were assumed to be independent of the features in the other group when conditioned upon class.

Connolly et al. tested the algorithms on a corpus of 80 news-agency articles with approximately 35,000 words in total. The training and testing data were generated in the following fashion. Correct candidates for each anaphor were paired with every other candidate for that anaphor and labeled as to which of the two antecedents was the correct candidate. To avoid teaching the classifier that one position of the pair is to be preferred, each candidate was presented to the classifier twice, with the candidate's position exchanged. The candidates themselves were presented as feature-value vectors, using the features in Table 4.7.[4]

[3] A similar method was recently proposed by Iida *et al.* (2003); I will not review their approach here, as it is concerned with resolving references of Japanese zero pronouns.

[4] With respect to recency, "zero" means same sentence as the anaphor; "one" means the previous sentence within the same paragraph; "near" means two or three sentences away or the previous sentence in a different paragraph; "far" means further afield. The features AGREE-KR1 and AGREE-KR2 were determined with respect to the system's lexicon and knowledge representation. No further details are

Level	Feature	Description	Values
NP	ANAPHOR-TYPE	NP type	pronoun, definite NP
NP	CANDIDATE1-TYPE	NP type	pronoun, definite NP
NP	CANDIDATE2-TYPE	NP type	pronoun, definite NP
NP	ANAPHOR-GRAM	Grammatical case	subject, V object, PP object, other
NP	CANDIDATE1-GRAM	Grammatical case	subject, V object, PP object, other
NP	CANDIDATE2-GRAM	Grammatical case	subject, V object, PP object, other
Coref	RECENCY1	Distance from anaphor to candidate 1	zero, one, near, far
Coref	RECENCY2	Distance from anaphor to candidate 2	zero, one, near, far
Coref	MORE-RECENT	Whether candidate 1 is more recent than candidate 2	
Coref	AGREE-COUNT1	Whether anaphor and candidate 1 agree in count	
Coref	AGREE-COUNT2	Whether anaphor and candidate 2 agree in count	
Coref	AGREE-GENDER1	Whether anaphor and candidate 1 agree in gender	
Coref	AGREE-GENDER2	Whether anaphor and candidate 2 agree in gender	
Coref	AGREE-KR1	Whether the anaphor meaning subsumes that of candidate 1	
Coref	AGREE-KR2	Whether the anaphor meaning subsumes that of candidate 2	

Table 4.7: Features used by Connolly *et al.* (1997).

Algorithm	Pronouns	Definite NPs	All references
Hand-crafted	51.6	25.7	38.7
Posterior classifier	32.7	7.9	20.4
Bayesian	35.9	11.2	23.7
Hybrid Bayesian	52.5	25.2	40.0
Decision tree	49.3	31.8	40.6
Hybrid decision tree	51.6	30.4	41.1
Neural net	52.1	28.9	40.6
Hybrid neural nets	**55.3**	**37.4**	**46.4**

Table 4.8: Performance of the classifiers used by Connolly *et al.* (1997) in percentage correct.

The classifiers's performance is shown in Table 4.8. The hybrid Neural Network classifier (subscape trained) achieved the best success rate of 37.4% for definite NP anaphors, 55.3% for pronominal anaphors, and 46.4% for all references.

4.3 A Machine Learning approach to other-anaphora

4.3.1 Experimental data

Five hundred samples of other-anaphors with NP antecedents were extracted from the *Wall Street Journal* corpus (Penn Treebank, release 2) using the data extraction procedure outlined in Chapter 3, with two exceptions. First, for the LEX and SAL experiments the anaphors were extracted in a two-sentence window; in the ML experiments reported in this chapter, I used a more realistic window of five sentences (sentence containing "(an)other" plus four preceding sentences). For anaphors that occur at the beginning of an article, where there may not be four preceding sentences, I used the context that was available. Second, since the LEX algorithm is not capable of handling pronominal antecedents, all pronouns were deleted from the LEX corpus prior to evaluation. For the ML experiments, pronouns were left intact.

available from the article.

The samples were extracted in three steps. First, I ran a TGREP2 search[5] to extract from all sections of the WSJ corpus other-NPs in a context of five sentences. The search returned 2,294 samples. Next, I used a regular expression grammar from Chapter 3 to extract samples with other-anaphors and non-structural antecedents.[6] During this procedure, 340 samples with idiomatic expressions, list-construction, etc. were discarded. In addition, 66 more such cases were discarded manually; these were the cases that the filter had missed. The resulting data set still contained samples of other-anaphors with non-NP antecedents. (Besides NPs, other-anaphors can take as antecedents, e.g. adjectives, clauses and various modifiers.) Samples with non-NP antecedents were filtered out during the corpus annotation stage, in which antecedents of other-anaphors were identified and assigned a tag ANTE or SPLANTE; the latter for split antecedents. There were at least 194 samples with non-NP antecedents in the corpus; I stopped annotation/manual filtering when I had 500 samples with non-structural NP antecedents. Also, six cases were such that the antecedents were cataphoric; these cases were also removed from the data set.

The resulting gold standard corpus subsumes the gold standard corpus used in Chapter 3. It has three times as many samples of other anaphors with NP antecedents as the gold standard corpus used in Chapter 3 (500 vs. 123 samples). Not all other-anaphors from Chapter 3, however, found their way to the new corpus. A few samples of the anaphors were filtered out during the corpus creation phaze. — The filters that filter out idiomatic phrases with “other”, reciprocal phrases and list-other are completely automatic; they were improved before running on the new corpus. This, however, resulted in the exclusion of some valid examples of other-anaphors. The new corpus is different from the one in Chapter 3 in one more respect. All anaphors were resolved from scratch in the new corpus, and in some cases (there were very few of them), this lead to a different interpretation of their antecedents. This is particularly true of cases in which more than one interpretation is possible.

As in the case with the corpora used for the pilot study of other-anaphora in the BNC (Chapter 2) and the comparison of the LEX and SAL algorithms on the samples

[5] `http://tedlab.mit.edu/~dr/Tgrep2/`

[6] This includes predicate nominals with “other”, e.g., “this is another company that . . . ”, and appositives, e.g., “Mary, the other teacher . . . ”.

from the *Wall Street Journal* corpus (Chapter 3), the gold standard for the machine learning experiments reported in this chapter was annotated by the author and it reflects my interpretation of what the antecedents of other-anaphors were in each particular case.

The antecedent extraction procedure was the same as in the LEX and SAL experiments in Chapter 3. The procedure is as follows. First, I extracted all base NPs, i.e., NPs that contain no further NPs within them. NPs containing a possessive NP modifier were split into a possessor phrase and a possessed entity phrase. Next, I filtered out null elements (tagged -NONE-). And finally, all anaphors and all potential and actual antecedents were lemmatised using the CELEX database.[7]

I used the following procedure (Soon *et al.*, 2001) to generate the training/test data set. Every pair of an anaphor and its closest preceding actual antecedent created a positive training instance. To generate negative training instances, I paired anaphors with each of the NPs that intervened between the anaphor and its antecedent. Note that this approach allows for only *one* positive instance per anaphor and thus it cannot handle split antecedents. In cases with split antecedents, the *closest* antecedent was paired with the anaphor to create a positive instance. Negative instances were generated as above, i.e., all instances prior to the most recent antecedent were discarded. This procedure produced a set of 3,084 antecedent-anaphor pairs, of which 500 (16%) were positive training instances.

4.3.2 The features

I experimented with twelve features automatically acquired from the corpus and from additional external resources. The features are summarized in Table 4.9. Eleven of these features were used in the first round of experiments, which compared a Naive Bayes classifier and a decision tree classifier. These features draw on previous research discussed in Section 4.2 and on my own work on other-anaphora described in Chapters 2 and 3. The last feature, WEB, was added in the second round of experiments, which compared two sources of semantic knowldge, WordNet and the Web. Below, I describe procedures for acquisition of features 1–11 and the resources I used for this purpose.

[7] http://www.kun.nl/celex/

No	Type	Feature	Description	Values
1	Gram	NP_FORM	Surface form (for all NPs)	definite, indefinite, demonstrative, pronoun, proper name, unknown
2	Gram	RESTR_SUBSTR	Does lemmatised antecedent string contain lemmatised anaphor string?	yes, no
3	Gram	GRAM_FUNC	Grammatical role (for all NPs)	subject, predicative NP, dative object, direct object, oblique, unknown
4	Gram	SYN_PAR	Anaphor-antecedent agreement with respect to grammatical function	yes, no
5	Gram	SDIST	Distance between antecedent and anaphor in sentences	1, 2, 3, 4, 5
6	Gram	MDIST	Distance between antecedent and anaphor in intervening NP units	$-N \ldots -1$
7	Sem	SEMCLASS	Semantic class (for all NPs)	person, organization, location, date, money, number, thing, abstract, unknown
8	Sem	SEMCLASS_AGR	Anaphor-antecedent agreement with respect to semantic class	yes, no, unknown
9	Sem	GENDER	Gender (for all NPs)	male, female, neuter, unknown
10	Sem	GENDER_AGR	Anaphor-antecedent agreement with respect to gender	same, compatible, incompatible, unknown
11	Sem	RELATION	Type of relation between anaphor and antecedent	same-predicate, hypernymy, meronymy, compatible, incompatible, unknown
12	Sem	WEB	Based on frequency counts for lexico-syntactic patterns on the WWW	webfirst, webrest

Table 4.9: Features used in the Naive Bayes and C4.5 classifiers. “Gram” stands for grammatical features, i.e., features that do not require semantic knowledge; “Sem” stands for semantic features, i.e., those that require (some) semantic knowledge.

The WEB feature will be introduced in Section 4.4.

Some features are attributes of a pair, e.g., RESTR_SUBSTR, SYN_PAR, SDIST, MDIST, SEMCLASS_AGR, GENDER_AGR, and RELATION. Other features are NP-level features, e.g., NP_FORM, GRAM_FUNC, SEMCLASS, GENDER, WEB. The features were computed for antecedent NPs only (all potential and actual antecedents).

4.3.2.1 Grammatical features

The feature NP_FORM is based on the POS tags in the WSJ corpus and heuristics. Generally, non-pronominal NPs were classified on the basis of their determiners and determinatives, based on (Huddleston and Pullum, 2002).[8] The values for this feature are summarized in Table 4.10. For instance, NPs with the determiners “the” and “all” are definite; NPs with the determinantives “every” and “several” indefinite[9]; and NPs with the determinantives “this” and “that” are demonstrative. The demonstrative category also includes pronouns “this”, ‘that”, “these” and “those”. Quantified phrases with cardinal numbers are indefinite; bare NPs are indefinite. Proper names are NPs with the start tag NNP or NNPS.[10] Proper names starting with “the”, e.g., “the Trade and Industry Ministry”, were classified as proper names. (Huddleston and Pullum, 2002) contains no information about definiteness of NPs with “certain”, “numerous”, and “various”; therefore they were classified as “unknown”.

The feature RESTR_SUBSTR matched lemmatized antecedent and anaphor strings and checked whether the antecedent string contains the anaphor string. This would apply to examples such as “one woman ringer ... another woman” and also “one woman ringer ... another ringer” and “one woman ringer ... another woman ringer”.

The values for the feature GRAM_FUNC were approximated from the parse trees and Penn Treebank annotation. In particular, clausal subjects, predicative NPs and

[8]A determiner is a function in NP structure; a determinative is a lexical category. Not all determiners are realized by determinatives, e.g., a genitive NP determiner. Also, many of the determinatives can have other functions than that of determiners, e.g., “three” is a determiner in “three books”, but a modifier in “these three books”.

[9]According to (Huddleston and Pullum, 2002), when using “each” and every”, the concern is with the individual entities (as opposed to a set, e.g., “both students”), but they do not satisfy the criterion of being identifiable.

[10]This is obviously an over-generalization, as not all NPs that start with the tag NNP are names, e.g., “Mitsubishi lawyers”.

Determiner or determinative	Value
“the”, “all”, “both”, “either”, “neither”, “no”, “none”	definite
“a(n)”, “each”, “every”, “some”, “any”, “(a) few”, “several”, “many”, “much”, “little”, “most”, “more”, “fewer”, “less”	indefinite
“this”, “that”, “these”, “those”	demonstrative
cardinal numbers	indefinite
bare NPs	indefinite
“certain”, “numerous”, “various”	unknown

Table 4.10: NP_FORM values.

some oblique adverbials are explicitly marked in the Treebank. Objects were inferred from the tree structures.

The feature SYN_PAR captured syntactic parallelism between anaphor and antecedent. If antecedent and anaphor had the same GRAM_FUNC value, they were syntactically parallel.

The features SDIST and MDIST measured the distance between anaphor and antecedent in terms of sentences and intervening NP units, respectively.

4.3.2.2 Semantic features

The SEMCLASS feature was determined as follows. Proper names were classified using ANNIE, part of the GATE2 software package.[11] Common nouns were looked up in WordNet, considering only their most frequent sense. To access the WordNet database, I used WordNet::QueryData Perl-to-WordNet interface.[12] In each case, the output was mapped onto one of the values in Table 4.9. The value “unknown” was used for concepts missing from WordNet.

The SEMCLASS_AGR feature compared the semantic class of the antecedent with that of the anaphor NP and returned “yes” if they belong to the same class; “no”, if they belong to different classes; and “unknown” if the semantic class of either the anaphor

[11] `http://gate.ac.uk`
[12] `http://www.ai.mit.edu/˜jrennie/WordNet/`

No	Feature	Nr of errors	% correct
1	NP_FORM	12	88%
2	RESTR_SUBSTR	0	100%
3	GRAM_FUNC	46	54%
4	SYN_PAR	16	84%
5	SDIST	0	100%
6	MDIST	0	100%
7	SEMCLASS	22	78%
8	SEMCLASS_AGR	20	80%
9	GENDER	18	82%
10	GENDER_AGR	18	82%
11	RELATION	22	78%

Table 4.11: Performance of feature acquisition modules on a random sample of 100 NPs.

or antecedent had not been determined.

The values for the feature GENDER were determined using lists of male and female titles, kinship and occupational terms, WordNet, and US Census 1990 lists of male and female first names.[13] Four values were used: “male”, “female”, and “unknown” for persons; and “neuter” for the rest of the nouns. The value “unknown” was used for nouns that can refer to both male and female persons, e.g., “teacher”.

The feature GENDER_AGR captured agreement in gender between anaphor and antecedent. Four values were possible: “same”, if both NPs have same gender; “compatible”, if antecedent and anaphor had compatible gender, e.g., “lawyer ... other women”; “incompatible”, e.g., “Mr. Johnson ... other women”; and “unknown”, if one of the NPs was undifferentiated, i.e., if the gender value was “unknown”.

The values for the feature RELATION (between other-anaphors and their antecedents) were partially determined by string comparison (“same_predicate”) and by looking up anaphor and antecedent synsets in WordNet (“hypernymy” and “meronymy”). An anaphor and antecedent stand in a same-predicate relation, if both NPs have the same head nouns, e.g., “one book ... other books”. Hypernymy is an ISA relation in Word-

[13] http://www.census.gov/genealogy/names/

Net, e.g., “the red car… other vehicles”. Meronymy is a part-of relation, e.g., “shares … other stocks”. As other relations, e.g., “redescription” (Examples 34–38 in Section 2.6), could not be determined on the basis of the information in WordNet, the following values were used: “compatible”, for NPs with compatible semantic classes, e.g., “woman … other leaders”; and “incompatible”, for NPs whose semantic classes were incompatible with each other, e.g., “woman … other economic indicators”. Two nouns are compatible if they have the same SEMCLASS value, e.g., “person”. The notion of compatibility is vague on purpose, because there is no explicit disjointness information in WordNet, except with respect to adjectives, which are disjoint with their antonyms, e.g., “big” is disjoint from “small”. Also, compatibility might be defined along a variety of parameters. For instance, if compatibility is based on the notion of being human, then nouns “woman” and “husband” would be compatible; if the compatibility is based on gender, then “woman” and “husband” would be incompatible. I calculated compatibility on the basis of values of the feature SEMCLASS. If the anaphor and antecedent head nouns had been tested for same_predicate, hypernymy, and meronymy relations, and no such relations were found to hold, then their SEMCLASS values were compared. If both nouns had the same value, e.g., “person”, then they were said to be of compatible type and the value of the feature RELATION was set to “compatible”. If the nouns had different SEMCLASS values, e.g., one of the nouns had the value “person” and the other had the value “date”, they were determined as incompatible. The value “unknown” was used if the type of relation could not have been determined, or when the SEMCLASS of one of the nouns was “unknown”.

All feature acquisition modules were implemented in Perl. Their performance is listed in Table 4.11, based on a random sample of 100 NPs. The performance of the modules was calculated based on a manual inspection of the values. For the feature SEMCLASS, I compared the real reading in the text with the most frequent sense in WordNet. It is unclear why the GRAM_FUNC module labelled correctly only 54% of the NPs in the random sample. The module tends to under-recognize oblique constituents and thus label them as “unknown”. Overall, errors made by the feature acquisition modules ranged from mistakes in conceptual design (e.g., the assumption that in the WSJ corpus the pronoun “they” is more likely to refer to persons), to errors made by the Named Entity Recognition module and omissions from WordNet.

Each antecedent NP in the training/test data set was represented as a feature-value vector, where the values described properties of antecedents. The first three entries in the data file on which the classifiers were trained are shown below. The first value is the NP identity tag; the last value is its class (i.e., with respect to antecedenthood). Note that the order of the features below is not the same as in Table 4.9. The values "webrest" and "webfirst" will be explained in Section 4.4.

```
NP1730,indef,abstract,yes,none,same,yes,same_pred,unknown,yes,2,-8,webfirst,yes
NP1731,def,date,no,none,same,no,diff,obq,no,2,-7,webrest,no
NP1732,def,org,no,none,same,no,diff,sbj,no,1,-6,webrest,no
```

4.3.3 Choosing the learning algorithm: Naive Bayes vs. decision trees

The type of learning framework is one of the three ingredients in any machine learning system. (The other two are the data and the features characterizing the data set.) Although most state-of-the-art coreference resolution systems use decision tree induction (Section 4.2), it is unknown whether decision trees are indeed the best approach for the task. And it is unknown whether decision trees would be the best approach for resolving other-anaphors, as the two phenomena are somewhat different. Specifically, not all definite NPs are anaphoric (actually less than half). And therefore, in resolving definite descriptions current coreference resolution systems tend to value higher approaches that result in higher precision than those that result in higher recall. Other-anaphors, on the other hand, *always* require an antecedent, and therefore it is important that as many of the anaphors that have been resolved have been resolved correctly.

I have experimented both with a decision tree classifier and a Naive Bayes classifier, using the implementations in the Weka machine learning library.[14] Both classifiers are implemented in Java; the C4.5 decision classifier is implemented as J48.[15]

[14] `http://www.cs.waikato.ac.nz/~ml/weka/`

[15] I have also experimented with a Neural Network classifier and Support Vector Machines with Sequential Minimal Optimization, SMO, both available from Weka. These classifiers achieved unsatisfactory performance on other-anaphors.

Both classifiers take as input a set of instances described by a conjunction of feature values (Section 4.3.2). On the basis of feature distribution in the training data set, they learn what features are more likely to predict a certain class. (With respect to other-anaphora, the classes are "antecedent=yes" or "antecedent=no".) Once the features and the best combination of them have been learned, the resulting model can be used to classify unseen instances.

The decision tree and the Naive Bayes classifiers differ with respect to the algorithms they employ. The decision tree learner constructs a binary decision tree, beginning with the question "which feature is the best predictor of a class c?". To answer this question, each feature is evaluated using a statistical test to determine how well it alone classifies the training examples. The best feature is selected and is used as the root node of the tree. A descendent node is created for each value of the root feature and the data set is then partitioned accordingly. The process is repeated at each descendent node in the tree, until there are no more instances to consider. At each node of the tree the learner selects a feature that partitions the samples in the most effective way. The most effective feature is the one with the highest *information gain*, i.e., the number of bits saved when encoding the target value of any arbitrary member of the data set (Mitchell, 1997). To avoid over-fitting, the decision tree produced by the classifier is automatically pruned to remove parts that are predicted to have a high error rate. The pruning confidence affects the way the error rates are estimated and hence the severity of pruning. Values smaller than the default (25%) cause more of the initial tree to be pruned; larger values result in less pruning. As decision trees can sometimes be difficult to understand, some implementations of C4.5 and its successor C5.0 provide a mechanism to convert trees into sets of rules. Rule sets are generally easier to understand than trees because each rule describes a specific context associated with the rule.

The Naive Bayes classifier uses Bayes Rule to estimate the probability of each class. The learning function is a product of the prior probability of a class c, $P(c)$, and the conditional probability of observing the feature values that support the class, $P(F|c)$:

$$P(c|F) = P(c)P(F|c)$$

Feature values are assumed to be independent; thus the probability of observing the conjunction of feature values is just the product of probabilities of individual features:

$$P(c|F) = P(c)\prod_i P(F_i|c)$$

The probabilities of individual feature values for each class are estimated by counting the frequencies with which a particular feature occurs in the data set among the positive and negative training samples (maximum likelihood estimation). The prior probability of a class is estimated in a similar fashion, by counting the number of positive and negative training instances in the corpus. Given two possible outcomes, the classifier chooses the class for which the learning function has the maximum value:

$$class = \underset{c\in\{yes,no\}}{\operatorname{argmax}} P(c)\prod_i P(F_i|c)$$

Because the resulting model is probabilistic, it can be difficult to interpret. However, I will show in Section 4.3.5.1 that there are ways to learn about the internal structure of the model, by interpreting the probabilities of feature values for each of the two classes.

The C4.5 classifier and the Naive Bayes classifier were trained and tested on features 1–11, using 10-fold cross validation. I experimented with different confidence values for the C4.5 classifier, from 25% to unpruned trees. The minimum number of instances per node for C4.5 was set at two.

The classifiers were evaluated using the following standard measures for ML algorithms. *Precision* indicates how many of the NPs classified as "antecedent=yes" are indeed the correct antecedents. *Recall* indicates how many of the NPs that should have been classified as "antecedent=yes" are correctly identified. *F-measure* is calculated as $F = 2PR/(R+P)$. For each classifier, I also give the number of *true positives* (TP) and *false positives* (FP). True positives are correct antecedents that have been successfully classified by the algorithm. False positives are NPs which have been classified as correct antecedents, while in fact they are not the target NPs. Performance of the Naive Bayes and C4.5 classifiers on other-anaphors is listed in Table 4.13.

To indicate the difficulty of the task, the performance of the classifiers is compared with a simple rule-based baseline algorithm `Base1`, which takes into account the ob-

Classifier	P	R	*F*	TP	FP
`Base1`	27.8%	27.8	27.8	139	361
C4.5 -C 25%	**69.3**	24.8	36.5	124	**55**
C4.5 -C 65%	54.2	33.2	41.2	166	140
C4.5 Unpruned	47.9	34.8	40.3	174	189
Naive Bayes	50.5	**40.6**	**45.0**	**203**	199

Table 4.12: Results for the Naive Bayes and the C4.5 classifiers trained on features 1–11 and comparison with the `Base1` algorithm. Bold font indicates best results. With respect to TPs and FPs, the best results are the highest number of TPs and the lowest number of FPs.

servation that the lemmatised head of an other-anaphor is often the same as that of its antecedent, as in Example 88.

(88) *These three countries* aren't completely off the hook, though. They will remain on a lower-priority list that includes **other countries** ...

For each anaphor, `Base1` string-compared its last (lemmatised) word with the last (lemmatised) word of each of its possible antecedents. If the words matched, the corresponding antecedent was chosen as the correct one. If several antecedents produced a match, the baseline chose the most recent one among them. If string-comparison returned no antecedent, the baseline algorithm chose the antecedent closest to the anaphor among all antecedents. The baseline assigned "yes" to exactly one antecedent per anaphor. Its precision, recall, and *F*-measure were 27.8%.

The C4.5 classifier achieved worse results overall than the Naive Bayes classifier induced from the same data using the same set of features. Its precision was significantly higher than the precision of the Naive Bayes classifier (for confidence value 25% only), but recall was much lower. (Recall was significantly lower for confidence values 25% and 65%. The difference between the unpruned C4.5 decision tree and Naive Bayes was not significant.)[16] Note that for the task of resolution of other-anaphora, recall is as important as precision. In terms of *F*-measure, the overall performance

[16]I used a t-test with confidence level .05 for all significance tests. In tables, significantly better precision, recall, and *F*-measure are in bold font.

of the Naive Bayes classifier was higher than that of the decision tree classifier, 45% vs. 41.2% (for confidence set at 65%; from now on all references to the decision tree classifier are to this version). Also, the Naive Bayes classifier consistently found more correct antecedents than the decision tree classifier: 203 TPs vs. 166 TPs for the C4.5, which amounted to a 22% difference in the scores. This was one of the reasons to abandon C4.5 and concentrate on Naive Bayes in subsequent experiments.

The second reason was the overall low recall scores for the C4.5 classifier. Decision tree classifiers tend to have low recall (and a relatively high precision), which is a result of the classifier's inductive bias: When partitioning a data set, a decision tree classifier orders features by an information-theoretic criterion: features that are expected to produce a larger reduction in entropy after the set has been partitioned according to this feature are applied sooner than those that result in a smaller reduction of entropy. This, in combination with post-pruning, leads to higher accuracy, but at the expense of discarding some features altogether. The Naive Bayes classifier, on the other hand, takes into account probabilities of *all* features for each particular instance. Moreover, it looks at the *combination of features* and this seems to lead to better results, in particular to a gain in recall, than a decision tree approach, even without post-pruning.

The third reason was the general unsuitability of the decision tree classifier for the data set. Decision tree classifiers are more suitable for data sets with many input dimensions (i.e., features), some of which might be redundant and/or noisy, and thus one of the classifier's tasks is to identify which features are relevant and informative with respect to the data set.[17] In the case of other-anaphors, I used relatively few, carefully chosen features, which were known in advance to be relevant for other-anaphors. It seemed wasteful to discard relevant features, even if they are not high-precision features.

And finally, as I indicated earlier, in the task of resolution of other-anaphora, a classifier's recall and precision are of equal importance, since (unlike definite descriptions) all other-anaphors require antecedents.

With these considerations, in all subsequent work I used only the Naive Bayes classifier, specifically the version trained on features 1–5,7,8,10, and 11, which from

[17] Richard Zemel, University of Toronto, p.c.

Classifier	P	R	*F*	TP	FP
Naive Bayes	50.5	40.6	45.0	203	199
`NBStand`	**51.7**	40.6	**45.5**	203	**190**

Table 4.13: Results for the Naive Bayes classifier trained on features 1–11 and `NBStand` trained on features 1–5,7,8,10,11.

now on is referred to as `NBStand`. Excluding features GENDER and MDIST lead to a small improvement in the classifier's precision and F-measure; `NBStand` made fewer false predictions, 190 FPs vs. 199 (Table 4.13).

In this section, I compared two learning frameworks on the task of resolution of antecedents of other-anaphors: decision trees (C4.5), which are commonly used in coreference resolution, and the Naive Bayes classifier. Two other classifiers, a Neural Network classifier and Support Vector Machines, were also tested on the data and abandoned, as they achieved unsatisfactory results. In future work, I would like to experiment with a Maximum Entropy classifier. Maximum entropy modeling has been used for a variety of natural language applications, from tagging (Ratnaparkhi, 1996) and PP phrase attachment (Ratnaparkhi, 1998), to named entity recognition (Borthwick *et al.*, 1998) and coreference resolution (Kehler, 1997b). Like decision trees and the Naive Bayes approaches, Maximum Entropy models recast the learning task as a probabilistic process. There is, however, one important difference: Maximum Entropy models make no assumptions about the probability distribution and thus they do not impose any additional constraints on the learned model. The Naive Bayes classifier, for instance, assumes that feature values characterizing instances are independent of each other, which is not true for all features; I will return to this issue in Section 4.3.5.2. Inductive bias (prior assumptions) is a fundamental property of inductive inference such as decision tree induction or Bayesian classification. A learner that makes no a priori assumptions about the learned classes has no rational basis for classifying unseen instances. However, it is desirable to keep the bias at a minimum and not assume anything beyond what has been observed in the data.

Despite its independence assumption, the Naive Bayes learner achieved encouraging results, and as I show in Section 4.3.5, this inductive bias did not seem to confuse the classifier. The next section examines in detail the classifier's performance, its limitations and learned model, and evaluates the contribution of the various knowledge sources.

4.3.4 `NBStand`: Performance, error analysis, and venues for improvement

4.3.4.1 Disambiguating the results

The results for the Naive Bayes classifier in Section 4.3.3 (and the decision tree classifier as well) are somewhat misleading: The classifier's performance is calculated without taking into account case boundaries. The classifier operates on the whole set of antecedent-anaphor pairs, and therefore for each anaphor it can classify as "yes" more than one potential antecedent. To amend this, some sort of incremental procedure is necessary. (Note that predicting multiple antecedents would not be a problem in coreference resolution, since antecedents often form coreference chains, and it would be desirable to identify all members of a coreference chain.) Also, the classifier does not know that each NP with "other" is anaphoric and therefore has an antecedent. (This again contrasts with definite NPs, which may be discourse-new and thus they would not require an antecedent.) Using a back-off procedure would remedy this problem, and I will describe such a procedure in Chapter 5. In this section, I focus on the first problem — that of multiple predictions per each anaphor case. I present a disambiguation procedure that takes into account the fact that each other-anaphor in the training and test data sets has only one antecedent. The procedure is as follows.

For each instance in the training/test data set, the `NBStand` classifier returned two posterior probabilities, the probability of "antecedent=yes" and the probability of "antecedent=no", which indicated the likelihood of a noun phrase to serve as antecedent of the other-anaphor. I assumed that NPs with class value above or equal 0.5 were likely to be antecedents of other-anaphors and those below 0.5 were not. For each anaphor in the training/test set, I considered each of the antecedent candidates in a right-to-left manner, starting from the anaphor NP. The procedure was incremental; it terminated when one antecedent was found (it may not have been the correct one) or

Classifier	Success rate	P	R	*F*	Correct	Incorrect	None
NBStand	**40.6**	51.7	**40.6**	**45.5**	n/a	n/a	n/a
$\texttt{NBStand}_D$	31.8%	**60.0**	31.8	41.57	159	106	235

Table 4.14: Performance of the NBStand classifier with and without the disambiguation procedure. The classifier that makes use of the disambiguation procedure is marked by a subscript $_D$.

when all NPs in the sample had been considered. Applying this procedure to the data set produced the following results: 159 samples in which the predicted antecedent was the correct one; 106 samples in which the predicted antecedent was not the correct one; and 235 samples in which no antecedent was predicted. Recast in terms of precision, recall and F-measure, these results were as in Table 4.14. For algorithms that make use of this disambiguation procedure I also used a *success rate* measure. The success rate indicates the percentage of anaphors with antecedents resolved to the correct NP.

The classifiers' performance is quite different with and without the disambiguation procedure. In particular, success rate and recall are 8.8 percentage points lower for $\texttt{NBStand}_D$ than for the algorithm without the disambiguation. Precision is however 8.3 percentage points higher. The new F-measure is somewhat lower, 41.57% vs. 45.5%. Note also the difference between TPs and correctly resolved cases on the one hand and FPs and incorrectly resolved cases on the other. As mentioned above, the classifier can classify more than one NP in each sample as antecedent. This is because the samples contain many *distractors*, i.e., entities that have antecedent properties but which are not the correct antecedents:

(89) Industrywide, oil production in this country fell by *500,000 barrels a day* to 7.7 million barrels in the first eight months of this year. Daily output is expected to decline by at least **another 500,000 barrels** next year.

The correct antecedent in Example 89 is the NP “500,000 barrels a day”; however, note the distractor “7.7 million barrels in the first eight months of this year” which has the same head noun as the anaphor and the correct antecedent. Distractors do not always have the same head noun as the anaphor NP. They can be names and common nouns with a different head:

(90) Delmed said yesterday that *Fresenius USA* would begin distributing the product and that the company is investigating **other possible distribution channels**.

In Example 90, the antecedent is the NP "Fresenius USA", which is related to the anaphor through redescription ("Fresenius USA is a possible distribution channel"). There is, however, another company name in this example, "Delmed", later referred to as "the company".[18] The approach I have outlined so far does not yet make any special provision for samples with redescriptions, and further research is necessary in this area. However, even if the algorithm were capable of modeling the redescription relation, it might not have been able to resolve this example after all because of the competition from the NP "Delmed". (Note that sentence level predication "would begin distributing the product" helps to disambiguate the antecedent.) I will return to the issue of redescription and other semantic relations not covered by the Naive Bayes algorithm in the next section.

In the next two sections I will give a detailed analysis of the types of errors made by the $\texttt{NBStand}_D$ classifier and make suggestions about how to improve the algorithm's performance. Section 4.3.5 will offer insights into the internal structure of the Naive Bayes model.

4.3.4.2 Error analysis of samples with zero antecedents

The $\texttt{NBStand}_D$ classifier failed to predict antecedents in 47% of cases for a variety of reasons. Some errors were the result of bugs in the preprocessing modules and noise in the data. Bugs ranged from extracting the wrong antecedent to errors in determining the gender or semantic class of antecedent, its relation to the anaphor, and in the string-matching module. Also, a better WordNet lookup procedure is required.

Word sense ambiguity was another major reason for missing antecedents. Antecedents were looked up in WordNet for their most frequent sense (the first sense in WordNet). However, there were at least 30 cases in the data sample such that either the anaphor or antecedent or both NPs were used in a sense that is ranked as less frequent

[18]It is possible that "the company" corefers with "Fresenius USA", however, this interpretation is less likely.

in WordNet. This might even be a quite frequent sense for a specific corpus, e.g., the word "issue" in the sense of "shares, stocks" in the WSJ. Therefore, there is a strong interaction between word sense disambiguation and resolution of other-anaphora. (See also (Preiss, 2002) for a similar claim for pronoun resolution.)

A better named entity resolution system would also improve the results. At least 20 cases were unresolved because their NE antecedents were misclassified.

Modifiers of various kinds contribute an important piece of meaning to the interpretation of anaphors and antecedents, e.g., the meaning of the noun phrase "a public/JJ figure/NN" is quite different from that of "a figure". The data I used in the ML experiments have been stripped of modification for obvious reasons. However, keeping the modifier "public" in the example above would have increased the chances of a correct resolution, since "public figure" is a separate entry in WordNet (and also sense 5 of the noun "figure"). Other modifiers are domain-specific, e.g., "cross-connect/JJ systems/NNS" and "telecommunications/NNS equipment/NN"; such adjective-noun and noun-noun collocations will not be recorded in WordNet. Furthermore, in cases such as Example 91, the antecedent should be interpreted against the PP complement of the anaphor, rather than its head noun:

(91) The big brokerage houses learned the art of the instant commercial after the 1987 crash, when they turned out *reassuring ads inviting investors right back into the stock market*. They trotted out **another crop of instant commercials** after the sudden market dip a few weeks ago.

In the example above, the phrase "another crop of instant commercials" means "instant commercials other than reassuring ads ..." and not "crop other than reassuring ads ...". In Section 3.6.1, I gave examples similar to Example 91 above; I referred to them as anaphors with semantically vague antecedents, e.g., "other types of watches". Example 91 is both similar to cases with vague anaphors and different: in the example above, the head of the anaphor is a quantity term, like "bushel", "flock" and "(a) couple". In general, it is not clear a priori whether antecedent NPs should be interpreted with respect to the head noun of the anaphor or with respect to its PP complement. Consider for instance the following pair, "New York ... other areas of the country" vs. "genetic engineering ... another area of promising research". While in the former example, the anaphor is interpreted as "areas (of the country) other than New York", in the latter, the anaphor is interpreted as "(promising) research other than genetic engineering". In the latter example, the noun "areas" is used for reasons of individuation, as in separating a particular amount and putting it in an individual container with mass nouns, e.g., " a cup of water". Further research is necessary to address the issue of modification in anaphor resolution.

Many samples of other-anaphors require knowledge that is not available from WordNet; I will return to this issue shortly. However, even if the information is available in WordNet, it might not be straightforward to retrieve, e.g., in Example 92 below, the antecedent "retinoblastoma" and anaphor "other cancers" are sister concepts; in Example 93, the antecedent "the machine" is a more general concept than the anaphor "computers" (i.e., they stand in an inverted relation to each other, as usually other-anaphors express more general concepts than their antecedents).

(92) "I was convinced that what was true of *retinoblastoma* would be true for all cancers." It was an audacious claim. But in Baltimore, Dr. Vogelstein, a young molecular biologist at Johns Hopkins Medical School, believed Dr. Knudson was right, and set out to repeat the Cavenee experiment in cells from **other cancers**.

(93) The computer can process 13.3 million calculations called floating-point operations every second. *The machine* can run software written for **other Mips computers** . . .

Consider also the following example:

(94) Dow Jones publishes *The Wall Street Journal, Barron's magazine,* **other periodicals and community newspapers** and operates electronic business information services .

The knowledge that a magazine is a periodical is reflected in WordNet, but only in the concept's gloss ("a periodic paperback publication"; for comparison, the concept "journal" has a hypernym "periodical"). Efforts have been made to include more and different types of links in WordNet for the purpose of reference resolution (Harabagiu, 1998; Harabagiu and Maiorano, 1999), but the resource that has been developed is not available publically and it was costly to produce.

The remainder of the samples can roughly be classified into five partially overlapping groups:

- Examples that require domain- or situation-specific knowledge or general world knowledge that is not available from WordNet, e.g., that a hurricane is a (natural) disaster; that a government can be an export customer; that steel is a commodity and coffee is an (important Colombian) export; that the precious metals sector is one of Dow Jones industry groups; that a business offer or proposal is a (business) transaction; and that being a customer means being in a business relationship with the services provider. See also Example 95:

  (95) One may be William Broderick, *a Sterling, Mass., grower.* "This is beautiful stuff, " he says, looking ruefully at big boxes of just-picked Red Delicious next to his barn." But I'm going to lose $50,000 to $60,000 on it. I'm going to have to get **another job** this year just to eat. "

- Examples involving bridging phenomena, sometimes triggered by a metonymic or metaphoric antecedent or anaphor, e.g., "The Justice Department's view ... other lawyers"; "chief executives ... other market sources"; "China General Plastic ... other investors". Consider also the following, rather striking example of bridging:

  (96) A Genentech spokeswoman said the agreement calls for Hoechst to promote TPA for *heart patients* and streptokinase for **other clot-reducing purposes**.

  The $\texttt{NBStand}_D$ classifier did resolve some metonymies, e.g., Example 97, perhaps through a combination of semantic and grammatical knowledge. But there were many metonymies in the data set that it could not handle.

  (97) *Unisys* dropped 3/4 to 16 1/4 after posting a third-quarter loss of $4.25 a share, including restructuring charges, but **other important technology issues** were mixed.

- Redescriptions and paraphrases, sometimes involving semantically vague anaphors and/or antecedents, e.g., “a question of investors’ access to the U.S. and Japanese markets ... other important economic issues”; “employment report ... other economic indicators”; “researchers ... two other research teams”; see also Example 90 above.

- Samples with ellipsis, e.g.,

  (98) He sees *flashy sports* as the only way the last-place network can cut through the clutter of cable and VCRs, grab millions of new viewers and tell them about **other shows** premiering a few weeks later.

  The antecedent in this example is not the flashy sports, but rather flashy sport shows or programs, and thus an important piece of information which is necessary to resolve the anaphor is omitted. (Alternatively, the antecedent is a `content-for-container` metonymy.[19]

- Samples with *collective* references, e.g., “pilots ... other labor group”; “Messrs. Cray and Barnum ... other senior management”; “us ... other firms”.

When I embarked on the machine learning approach to other-anaphora, I hypothesized that a probabilistic resolution algorithm that takes into account a variety of semantic and grammatical factors might be able to resolve more samples of other-anaphors with metonymic and semantically vague antecedents and redescription cases than an approach that primarily relies on semantic knowledge. While this turned out to be true — to some extent: the Naive Bayes classifier resolved a handful of metonymies and bridging antecedents — there are many such samples in the data set which it could not resolve. This confirmed the importance of semantic and in particular domain-specific and general world knowledge for resolution of other-anaphors. It further highlighted the need for further research into the nature of bridging references and the role of modification in anaphora resolution.

[19]While Example 98 might be an example of bad writing, as has been suggested by several people, such examples occur in natural texts and a robust anaphor resolution system should be able to handle them. Also, if using some sort of filtering to filter out such examples, where would one draw a line between “bad writing” and “good writing” perhaps expressing a controversial opinion, as, e.g., in Example 38? Neither of these examples can be resolved using a conventional knowledge base.

4.3.4.3 Error analysis of samples with incorrect predictions

Two questions require an answer with respect to cases of other-anaphors in which the classifier made incorrect predictions: (1) why wasn't the correct antecedent predicted; and (2) why did the classifier choose incorrect entities? To answer these questions, I modified the disambiguation procedure to consider *all* NPs in each other-anaphor sample, instead of terminating it as soon as one antecedent was found. The results fell into two groups: (1) cases in which one antecedent or more were predicted and none of them was correct, and (2) cases in which more than one antecedent were predicted with one of them being the correct one. On average, in both groups, the algorithm predicted 1.5 antecedents per each anaphor. As for the first question posed at the beginning of this section, the algorithm did not find correct antecedents for the same reasons as in Section 4.3.4.2, e.g., bugs in preprocessing modules, word sense ambiguities, incorrect NE classifications, lack of domain, situation-specific or general knowledge and metonymic and redescription relations. It made incorrect predictions for the following reasons:

- There is a bias towards named entities. Because there are so many proper names in the WSJ corpus and because almost 40% of the correct antecedents are proper names, named entities are more likely to be predicted as antecedents. (See also Section 4.3.5.1.)

- NPs with the same head noun as the anaphor are also more likely as antecedents than NPs with a different head (Section 4.3.5.1).

- A syntactic filter is necessary to filter out impossible antecedents on the basis of syntactic constraints reported in Chapter 3.

- Some correct antecedents just can't make it over the threshold, their "antecedent=yes" probability hovering around 0.48. This is especially true of pronominal antecedents, abstract and "thing" entities, and antecedents whose grammatical function in the sentence is oblique or unknown. (See also Section 4.3.5.1.)

There are, however, genuinely hard cases, such as Example 89 in the previous section and Example 99 below:

(99) But Coleco bounced back with the introduction of the *Cabbage Patch dolls*, whose sales hit $600 million in 1985. But as the craze died, Coleco failed to come up with **another winner** and filed for bankruptcy-law protection in July 1988.

The correct antecedent in Example 99 is the NP "Cabbage Patch dolls". The algorithm resolved the antecedent to "Coleco", because it is a name and subject of the sentence (the NER module failed to classify it as an organization). Note also that both NPs stand in a metonymic relation to the anaphor, and as such they are both likely as antecedents of "another winner".

4.3.5 Explaining the errors

4.3.5.1 An insight into the structure of the Naive Bayes model

To fully understand the reasons why the `NBStand` classifier failed to find antecedents in 68.2% of samples of other-anaphors, it is necessary to examine the Naive Bayes model at a micro level, by comparing the conditional probabilities of each feature value for each of the two classes. Recall from Section 4.3.3 that the classifier predicts the class of each new instance on the basis of the prior probability of each of the two classes and the conditional probabilities of each feature value given class:

$$class = \underset{c \in \{yes,no\}}{\text{argmax}} \; P(c) \prod_i P(F_i|c)$$

The prior probabilities of "c=yes" and "c=no" are constant for the data set; $P(c = yes)$ is 0.16 and $P(c = no)$ is 0.84. The conditional probabilities, on the other hand, range from 0.0012 for $P(relation = holonym|c = no)$ to 0.9861 for $P(restr\_substr = no|c = no)$. To learn which features, in particular which feature values, are good predictors of a class c, consider the ratio

$$P(v_i|c)/P(v_i|\overline{c}),$$

where $\overline{c}$ is the complement class of c and v_i is some feature value. The greater the ratio, the more likely the value of a particular feature to predict that class: For some instance i, if *most* of the features have values such that $P(value|c = yes)$ is much greater

FEATURE=value	Class	Ratio
RELATION=same_pred	yes	13.23
RESTR_SUBSTR=yes	yes	12.18
GENDER_AGR=same	yes	2.63
SEMCLASS_AGR=yes	yes	2.25
NP_FORM=name	yes	2.05
RELATION=incomp	no	2.03
GRAM_FUNC=subject	yes	1.75
SEMCLASS=org	yes	1.78
SEMCLASS_AGR=no	no	1.73
GENDER_AGR=incomp	no	1.54
NP_FORM=indef	no	1.52
RELATION=unknown	yes	1.51

Table 4.15: Values that are most likely to predict a particular class, when aggregated. Values with ratio below 1.5 are not included.

than $P(value|c = no)$, then, when aggregated, their product will be high enough to counteract the low prior for "c=yes" and, as a result, classify the instance as "c=yes". The reverse also holds, though with more force: For some instance *i*, if *many* of the features have values such that $P(value|c = no)$ is much greater than $P(value|c = yes)$, then the instance is more likely to be classified as "c=no". It doesn't take many features to tip the scale one way or the other; in Example 99 in the previous section, it sufficed that the NP "Coleco" was a proper name and subject of the sentence to classify it as antecedent (with *P=0.67*). The values with the greatest $P(v_i|c)/P(v_i|\overline{c})$ ratio are listed in Table 4.15 (excluding infrequent values such as "SEMCLASS=numeric", and "GRAM_FUNC=predicate".)

In general, a noun phrase is more likely to be classified as antecedent if:

- it has the same noun head as the anaphor;
- it agrees with the anaphor in gender and/or semantic class;
- it is a named entity;

- it is a clausal subject;
- it is a name of an organization;
- the relation between the anaphor and antecedent is unknown.[20]

A noun phrase is more likely to be classified as non-antecedent if:

- the anaphor and antecedent are of incompatible semantic class;
- the anaphor and antecedent do not agree with each other in their semantic class value and/or gender;
- if the antecedent is realized as an indefinite NP.

Semantic values dominate Table 4.15, confirming the claim above and in Chapter 3 that semantics is an important factor for resolving the antecedents of other-anaphors.

4.3.5.2 The contribution of the features

I showed in Section 4.3.4 that the `NBStand` classifier was mostly unsuccessful in resolving other-anaphors with semantically vague, metonymic and other bridging antecedents and the redescription cases. Perhaps adding grammar wasn't necessary at all? To evaluate the relative contribution of the various knowledge sources to the classifier's performance, I ran a series of leave-one-out classifiers, where I disabled one feature at a time (features 1–11). (The disambiguation procedures was not used.) This exercise also shed some light on the interaction of the features.

The Naive Bayes classifier makes an assumption that the features it uses are conditionally independent of each other. However, it is clear that many of them are not. For instance, the feature SEMCLASS_AGR is dependent on SEMCLASS, because it makes use of the semantic class information; the feature GENDER is dependent on NP_FORM and SEMCLASS; and the feature SYN_PAR is based on the values for the feature GRAM_FUNC. These dependencies might reduce the power of the Naive Bayes to

[20] This should not be surprising. There are many samples of other-anaphors which are redescriptions and which involve bridging and metonymic antecedents; these types of antecedents are currently assigned value "unknown".

Feature	P	R	*F*	TP	FP
NP_FORM	**52.0**	35.8	42.4	179	165
SEMCLASS	**52.3**	36.6	43.1	183	167
SEMCLASS_AGR	**50.6**	34.6	41.1	173	169
GENDER	**51.8**	40.4	**45.4**	202	188
GENDER_AGR	**53.8**	37.2	44.0	186	160
RESTR_SUBSTR	47.6	**41.4**	44.3	207	228
RELATION	**53.3**	32.4	40.3	162	142
GRAM_FUNC	50.1	37.6	43.0	188	187
SYN_PAR	49.3	39.8	44.0	199	205
SDIST	50.4	40.4	44.8	202	199
MDIST	**50.9**	**40.6**	**45.2**	203	196

Table 4.16: Results for leave-one-out classifiers (features 1–11). Text in bold font indicates success rate, precision, recall, and F-measure equal to or better than those when using a full feature set.

discern what is going on, and hence removing these dependencies is likely to lead to improved results.

When evaluating the performance of leave-one-out classifiers, I assumed that if the algorithm's performance did not change when a particular feature was disabled, then the feature did not make a significant contribution. If the performance of the classifier increased, the feature had a confusing effect and it should be removed. With the majority of the features, recall dropped 0.2–8.2%, while precision varied between 47.6% and 53.8% (Table 4.16). Moreover, for many of the features, recall dropped and precision (naturally) increased. However, three features equally affected precision and recall: GRAM_FUNC, SYN_PAR, and SDIST. When these features were disabled, both precision and recall dropped. For two features, GENDER and MDIST, the F-measure actually increased, suggesting that these features might be redundant. In fact, when GENDER and MDIST were disabled, the classifier achieved a somewhat higher performance, in particular a higher precision (Table 4.13).

In addition to leave-one-out classifiers, I ran a series of one-feature classifiers, e.g.,

Features	P	R	F	TP	FP
RELATION	68.0	16.6	26.7	n/a	n/a
RESTR_SUBSTR	**70.6**	**16.8**	**27.1**	n/a	n/a

Table 4.17: Results for one-feature classifiers with non-zero F-measure.

Features	P	R	F	TP	FP
RELATION & RESTR_SUBSTR	68.5	17.0	27.2	n/a	n/a

Table 4.18: Results for the two-feature classifier (features RELATION and RESTR_SUBSTR).

the classifier was trained and tested on one feature at a time. Two features gave non-zero F-measure: RELATION and RESTR_SUBSTR (Table 4.17). The RELATION feature is not independent of RESTR_SUBSTR; one of the values of the feature RELATION is "same_predicate", which is based on the output of the RESTR_SUBSTR module. However, this interdependency did not seem to confuse the Naive Bayes classifier; when trained on just these two features, the classifier actually achieves slightly better results than one-feature classifiers (Table 4.18).

To evaluate the joint contribution of similar knowledge sources, e.g., grammatical and semantic features, I ran two additional baseline classifiers, `NBBaseGR` (grammatical features only), and `NBBaseSEM` (semantic features only).

Results in Table 4.19 confirm that grammar by itself is not sufficient to resolve other-anaphors. In particular, recall of the `NBBaseGR` classifier is very low, 14.6%. In fact, `NBBaseGR` performed worse than `Base1`, which operated on a combination of recency and string matching. This difference is due to the inherent limitation in the Naive Bayes approach: The training data contains no indication about *how many* antecedents there are in the data set for each anaphor sample, while `Base1` always knows that for each anaphor in the data set there is one and only one antecedent (Section 4.3.4). Compared with `NBBaseGR` and `Base1`, the semantic baseline `NBBaseSEM` performed significantly better, achieving a recall of 34.8%. However, precision dropped significantly, to 49.3% (as compared with precision of `NBBaseGR`). Nevertheless, the performance of

Classifier	P	R	*F*	TP	FP
`Base1`	27.8	27.8	27.8	139	361
`NBBaseGR`	74.5	14.6	24.4	73	**25**
`NBBaseSEM`	49.3	34.8	40.8	174	179
`NBStand`	**51.7**	**40.6**	**45.5**	**203**	199

Table 4.19: Results for the `NBBaseGR` and `NBBaseSEM` classifiers and comparison with `Base1` and `NBStand`. `NBBaseGR` was trained on grammatical features 1–5; `NBBaseSEM` was trained on semantic features 7,8,10,11.

`NBBaseSEM` approached that of `NBStand` (the difference is insignificant).

Another interesting observation is the proportion of true and false positives: while `NBBaseGR` found three times more TPs than FPs, `NBBaseSEM` found slightly more FPs than TPs. Combining the two types of knowledge sources (as `NBStand`) seemed to have had a *tempering* effect on the semantics: the number of TPs for `NBStand` was higher than the number of FPs. Also, semantics and grammar together seemed after all to achieve better results than either of them by itself, suggesting that having more knowledge (of a different kind) is better when resolving antecedents of other-anaphors.

4.4 Naive Bayes with the Web: `NBStand+Web`

4.4.1 The method

While the performance of the `NBStand` classifier was encouraging, it was not yet satisfactory. I showed in Section 4.3.4.3 and in Chapter 3 that semantic knowledge is the most important source of information in determining the antecedents of other-anaphors. However, currently available semantic resources such as WordNet are insufficient for this task, because they often lack the kind of knowledge that is necessary to find the antecedents of other-anaphors. That WordNet is inadequate for the task of reference resolution has also been pointed out by, e.g., Vieira and Poesio (2000). Vieira and Poesio reported that in their corpus of bridging references, in almost 40% of cases, the relation between anaphor and antecedent was something other than hyponymy, synonymy, and meronymy.

There have been efforts to extract missing lexical relationships from corpora in order to build new knowledge sources and to enrich existing ones (Hearst, 1992; Berland and Charniak, 1999; Poesio *et al.*, 2002). However, the size of the corpora used still leads to data sparseness (Berland and Charniak, 1999) and the extraction procedure can therefore require extensive smoothing. A more promising venue of research has recently been demonstrated by Markert, Nissim, and Modjeska (2003), henceforce MMN. MNM presented a novel method for anaphora resolution which uses the Web as the primary source of semantic information. They searched the Web with lexico-semantic patterns specific to each anaphor type. The basic idea is simple: if an antecedent and anaphor are linked by a semantic relation which is only implicitly expressed, e.g., that in American English universities are informally called schools, as in Example 100, there are cases in which the same relation is expressed *explicitly* (i.e., in a conjunction), e.g., Example 101:

(100) As the session broke up, I was approached by a man identified himself as the alumni director of *a Big Ten university* "I'd love to see sports cut back and so would a lot of my counterparts at **other schools**, but everybody's afraid to make the first move, he confided.

(101) Foreign students obtain student visas from US consulates abroad after they are accepted by *US colleges, universities*, and **other schools**.

Specifically, in Example 101, the antecedent "US colleges, universities" is available structurally, as the left conjunct of the anaphor. In such constructions, which I have called list-constructions, the left conjunct is (almost) always the antecedent of the other-anaphor. There are several other constructions that structurally explicitly express a hyponymy, similarity or other relation between the lexical head of the anaphor and the antecedent, e.g., "X(s) such as Y(s)". (See (Hearst, 1992) for other patterns.)

MNM used these lexico-syntactic patterns to collect knowledge needed to resolve other-anaphors and bridging (meronymy) samples. Specifically, for other-anaphors, they used the list-construction "X(s) and other Y(s)". This pattern was instantiated for each antecedent-anaphor pair in their corpus, including actual and potential antecedents. In Example 100, the instantiations were, e.g., `director and other schools`, `university and other schools`, `sports and other schools`, etc. These instantiations were submitted as queries to the Google search engine, and the number of hits was counted. Their rationale is that the most frequent of these instantiations is a good clue for the antecedent. For instance, the query `universities and other schools` yielded over 700 hits, while the other two queries yielded under 10 hits each.

As documents can contain instantiations of the list-pattern with singular and plural antecedents, MNM used the following pattern:

$(\mathtt{N}_1\{sg\}$ `OR` $\mathtt{N}_1\{pl\})$ `and other` $\mathtt{N}_2\{pl\}$

where $\mathtt{N}_1$ and $\mathtt{N}_2$ are variables, to be substituted with the rightmost nouns of lexical heads of anaphors and antecedents and "OR" is the boolean operator. (MNM used only the rightmost nouns to avoid data sparseness.) The pattern above was instantiated differently for common noun and proper name antecedents. For common noun antecedents, MNM instantiated the pattern by substituting $\mathtt{N}_1$ with each possible antecedent and $\mathtt{N}_2$ with the anaphor. For proper name antecedents the pattern was $\mathtt{N}_1$ `and other` $\mathtt{N}_2$, e.g., `Mr. Pickens and other shareholders`. In addition to the proper name pattern, they used two additional instantiations of the common-noun pattern for samples with proper name antecedents: (1) Since using proper names in Web queries would lead to data sparseness, they substituted $\mathtt{N}_1$ with the antecedent's NE category and $\mathtt{N}_2$ with the anaphor head noun, e.g., `(person OR persons) and other shareholders`; and (2) they substituted $\mathtt{N}_1$ with the anaphor head noun and $\mathtt{N}_2$ with the antecedent's NE category, e.g., `(shareholder OR shareholders) and other persons`. Pattern (2) was necessary because in this example the antecedent NP "person" is not a hyponym of "shareholder" (in WordNet sense), but rather the anaphor "shareholder" is a hyponym of "person".

MMN generated such patterns for all anaphors in their corpus and submitted them as queries to the Google search engine making use of its API technology. For each pattern instantiation, they obtained its raw frequency on the Web. The frequency counts were scored, taking into account the individual frequencies of anaphors and antecedents by adapting Mutual Information (MI). For each sample of other-anaphor in their corpus, the noun phrase with the highest MI score was then proposed as antecedent of that anaphor. If two antecedents had achieved the same score, a recency based tie-breaker chose the antecedent closest to the anaphor.

In collaboration with Katja Markert and Malvina Nissim (Modjeska *et al.*, 2003), I obtained the Web MI scores for all NPs in my training corpus and incorporated them into the Naive Bayes classifier as a separate feature WEB in the following fashion. For each anaphor, the antecedent with the highest MI score got feature value "webfirst". (If several antecedents had the highest MI, they all got value "webfirst".) All other antecedents got the feature value "webrest". I chose this method of integrating the web score into the WEB feature instead of, e.g., giving score intervals, for two reasons. First, since score intervals are unique for each anaphor, it is impossible to incorporate them into a machine learning framework in a consistent manner. Second, this method introduces an element of competition between several antecedents (Connolly *et al.*, 1997), which the individual scores do not reflect.

4.4.1.1 Results

Adding the WEB feature significantly improved the classifier's performance. The classifier[21], referred to as `NBStand+Web`, achieved a 9.1 percentage point improvement in precision (an 18% improvement relative to the `NBStand` classifier) and a 12.8 percentage point improvement in recall (32% improvement relative to the `NBStand` classifier), which amounted to an 11.4 percentage point improvement in F-measure (25% improvement relative to the `NBStand` classifier); see Table 4.20.

When run with the disambiguation procedure (Section 4.3.4), the classifier's results were as follows: 213 correctly resolved cases, 138 cases with incorrect predictions, and 149 cases with zero-antecedents. Table 4.21 recasts these results in terms of success rate, precision, recall, and F-measure.

$\texttt{NBStand+Web}_D$ found significantly more correct antecedents than $\texttt{NBStand}_D$, achieving a 42.6% recall vs. 31.8%. In particular, it had a higher success rate with samples of other-anaphors that required domain-specific or general world knowledge. For instance, it correctly resolved the following cases from Section 4.3.4.2: "[hurricane] Hugo ... other disasters"; "steel ... other commodities"; "coffee ... another important

[21] Trained on features 1–5,7,8,10–12; the features GENDER and MDIST worsened the overall performance of the classifier slightly.

Classifier	P	R	*F*	TP	FP
`Base1`	27.8	27.8	27.8	139	361
`NBBaseGR`	**74.5**	14.6	24.4	73	**25**
`NBBaseSEM`	49.3	34.8	40.8	174	179
`NBStand`	50.5	40.6	45.0	203	199
`NBStand+Web`	60.8	**53.4**	**56.9**	**267**	172

Table 4.20: Results for the `NBStand+Web` classifier and comparison with `NBStand`, `Base1`, `NBBaseGR` and `NBBaseSEM` classifiers.

Classifier	Success rate	P	R	*F*	Correct	Incorrect	None
`NBStand`$_D$	31.8%	60.0	31.8	41.57	159	**106**	235
`NBStand+Web`$_D$	**42.6%**	**60.68**	**42.6**	**50.06**	**213**	138	**149**

Table 4.21: Performance of the `NBStand+Web` classifier with the disambiguation procedure and comparison with `NBStand`$_D$.

Colombian export"; "the precious metals sector ... other Dow Jones industry groups"; "[business] offer ... other transactions/alternatives"; and "the magazine ... other periodicals". It also resolved the following cases: "Romanee-Conti ... another Burgundy estate"; "Columbia ... other thrifts".

`NBStand+Web`$_D$ was more successful than `NBStand`$_D$ in handling metonymic and other bridging antecedents. For instance, it correctly resolved the following examples which `NBStand` did not: "Samsung ... other producers" and "Compaq Computer ... other technology issues".

`NBStand+Web`$_D$ handled well some redescriptions, paraphrases and vague anaphors, e.g., "pound concerns ... another boon for the dollar"; "a question of investors' access to the U.S. and Japanese markets ... other important economic issues"; "increasing costs as a result of greater financial exposure ... other, far-reaching repercussions"; "this court ruling ... other cases"[22]; "the decline ... other indicators"; and "this measure ... other economic indicators". There were, however, redescription examples that

[22]But not "a single court decision ... other, less compelling cases"

were so unusual and/or situation- or speaker-specific that even the Web would not be much of help, e.g., Examples 102 and 103:

(102) He liked *the well-lighted lobby display of Honda's cars and trucks* so much that he had Nissan's gloomy lobby exhibit refurbished. Later, Nissan borrowed **other Honda practices, including an engineering "idea contest" to promote inventiveness**.

(103) ... *recent strong growth in dividends* ... **another warning flag**

The ellipsis example in Section 4.3.4.2 "flashy sports ... other shows" (Example 98) and some examples with inverted relations between the anaphor and antecedent, e.g., "the machine ... other Mips computers" were also correctly resolved.

$\texttt{NBStand+Web}_D$ was immune to the problem of word sense disambiguation. Unless the patterns used to search the Web occur at a clause boundary (more about this below), we are *guaranteed* to have the correct word sense: within a specific lexico-semantic pattern, the anaphor and antecedent constrain each other's sense within the context of the pattern, e.g., the noun "bank" in "the bank and other financial institutions" can only mean a financial institution and not a bank of a river. However, like $\texttt{NBStand}_D$, $\texttt{NBStand+Web}_D$ was sensitive to NER failures; incorrect NE classifications lead to incorrect pattern instantiations and thus incorrect co-occurrence frequencies.

While the algorithm's recall was significantly higher than that of $\texttt{NBStand}_D$, its precision was almost identical to the precision of $\texttt{NBStand}_D$, 60.68% vs. 60.0%, i.e., the classifier found more antecedents, but they were not necessarily the correct ones. Note also that $\texttt{NBStand+Web}_D$ incorrectly resolved more other-anaphors than $\texttt{NBStand}_D$, 138 cases vs. 106 cases (Table 4.21). There are two reasons for this. First, because the Web lookup method does not postprocess the returned documents in any way (besides returning frequencies), it can not determine whether the observed pattern $\texttt{NP}_1$ $\texttt{and other NP}_2$ occurs within a noun phrase or at a clause boundary as in Example 104:

(104) Studies have been done that show *some people* do not want to live in the area because of a bridge and **other people who choose to live in the area** because [...]

This leads to incorrect frequency and mutual information scores, to incorrect instantiation of the WEB feature, and ultimately incorrect predictions.

Second, some collocations are more *natural* than others, e.g., the pattern `contract OR contracts and other issues` (304 occurrences) is more frequent than `interpretation OR interpretations and other issues` (47 occurrences). Even when the frequency scores were adjusted to take into account the individual frequencies of words "contract(s)", "interpretation(s)" and "issue(s)", the noun phrase with the head "contracts" achieved a higher MI score and thus was more like to be the antecedent of "other issues", while in fact in this example, the correct antecedent was the NP "interpretations".

4.4.2 WordNet semantics or the Web?

In the previous section, I cited many examples of other-anaphors that $\texttt{NBStand}_D$ could not resolve and which $\texttt{NBStand+Web}_D$ resolved correctly. Given this difference, is Web semantics better than WordNet semantics, or are the two resources complimentary? To answer this question, I trained the Naive Bayes classifier on the grammatical features 1–5 and the WEB feature, i.e., excluding the feature MDIST and all WordNet-based features (all gender, semantic class, and relation features). This classifier is referred to as `NBBaseGR+Web`. The classifier's performance from 10-fold cross-validation is listed in Table 4.22; with the disambiguation procedure, in Table 4.23. Its performance is compared with the performance of `NBBaseGR`, `NBBaseSEM`, `NBStand` and `NBStand+Web` classifiers with and without the disambiguation procedure.

It is clear from the tables that the Web contains more information relevant to resolution of other-anaphors: The `NBBaseGR+Web` classifier significantly outperformed `NBStand` in all measures. (`NBStand` used WordNet semantics.) Note also a significant drop in the number of false positives, from 199 to 128 instances. It is also interesting to note the difference in performance between the `NBBaseGR+Web` classifier which used only Web knowledge and `NBStand+Web` which used both Web and WordNet knowledge. At first sight (Table 4.22), `NBBaseGR+Web` achieved a slightly lower recall than `NBStand+Web` and a significantly higher precision. However, a comparison of the results with the disambiguation procedure (Table 4.23), shows that $\texttt{NBBaseGR+Web}_D$

Classifier	P	R	*F*	TP	FP
NBBaseGR	74.5	14.6	24.4	73	**25**
NBBaseSEM	49.3	34.8	40.8	174	179
NBStand	50.5	40.6	45.0	203	199
NBStand+Web	60.8	**53.4**	56.9	**267**	172
NBBaseGR+Web	**67.0**	52.0	**58.6**	260	128

Table 4.22: Results for NBBaseGR+Web trained on grammatical features 1–5 and WEB and comparison with NBStand+Web, NBStand, NBBaseGR and NBBaseSEM classifiers.

Classifier	Success rate	P	R	*F*	Correct	Incorrect	None
NBStand$_D$	31.8%	60.0	31.8	41.57	159	**106**	235
NBStand+Web$_D$	42.6%	60.68	42.6	50.06	213	138	149
NBBaseGR+Web$_D$	**49.2%**	**66.85**	**49.2**	**56.68**	**246**	122	**132**

Table 4.23: Performance of the NBBaseGR+Web$_D$ classifier with the disambiguation procedure and comparison with NBStand$_D$ and NBStand+Web$_D$ classifiers.

achieved both a higher recall, 49.2% vs. 42.6% (the difference is not significant), and a significantly higher precision, 66.85% vs. 60.68%, than NBStand+Web$_D$. This suggested that having more semantic knowledge (from different sources) available might not necessarily lead to better predictions. One possible explanation for this is that WordNet often provides the algorithms with wrong information, because no word sense disambiguation was performed when looking up words in WordNet. In the future, I would like to test this hypothesis by having an oracle which would always give the correct word sense.

If Web information is more relevant in resolving antecedents of other-anaphors, perhaps one could suspend all other information sources all together and rely just on the Web? To verify this hypothesis, I trained the Naive Bayes classifier on just the WEB feature; this classifier is referred to as NBJustWeb. Indeed, a first impression confirmed that using just the WEB feature would be sufficient to resolve other-anaphors (Table 4.24). However, notice the high number of false positives, 239 instances, or

Classifier	P	R	*F*	TP	FP
NBStand	51.7	40.6	45.5	203	190
NBStand+Web	60.8	53.4	56.9	267	172
NBBaseGR+Web	**67.0**	52.0	**58.6**	260	**128**
NBJustWeb	56.6	**62.4**	**59.4**	**312**	239

Table 4.24: Results for NBJustWeb, trained on just the WEB feature and comparison with NBBaseGR+Web, NBStand+Web, and NBStand classifiers.

47.8%. And once the classifier was tested with the disambiguation procedure (Table 4.25), it became clear that just using the Web information might not be sufficient after all. $\texttt{NBJustWeb}_D$ achieved a somewhat higher recall than $\texttt{NBBaseGR+Web}_D$, 52.8% vs. 49.2%. (The difference was not significant.) However, its precision was 13.08 percentage points below that of $\texttt{NBBaseGR+Web}_D$. (This difference was significant.) While $\texttt{NBJustWeb}_D$ attempted to resolve a lot more cases than $\texttt{NBBaseGR+Web}_D$ (only 9 samples produced zero antecedents), it got almost half of the cases incorrect, 227 instances, or 45.4%. (Of the 500 correct antecedents, 312 cases or 62.4% are "webfirst"; the remaining 188 cases are "webrest". And among the 2,584 negative training instances, 239 NPs have value "webfirst".) Here, as in the case of the $\texttt{NBBaseSEM}_D$ and $\texttt{NBStand}_D$ classifiers, grammar seems to have had a *tempering* effect on (the Web) semantics, which alone tends to over-generate. One can, however, take advantage of the WEB feature's tendency to better recall and resolve more samples of other-anaphors by using the WEB feature as a back-off procedure. I will support this claim in the next chapter, in which I will present a hybrid method for resolving other-anaphors.

4.5 A more realistic test

In the discussion so far, all classifiers were tested on data generated in the same fashion as the training data set, i.e., all NPs prior to the actual antecedent were removed from the data set. This introduced an undesirable bias: in a more realistic data set, a classifier could have classified more entities as "antecedent=yes", including those that appear in the text before (i.e., to the left of) the actual antecedent.

Classifier	Success rate	P	R	F	Correct	Incorrect	None
$\texttt{NBStand}_D$	31.8%	60.0	31.8	41.57	159	**106**	235
$\texttt{NBStand+Web}_D$	42.6%	60.68	42.6	50.06	213	138	149
$\texttt{NBBaseGR+Web}_D$	49.2%	**66.85**	49.2	**56.68**	246	122	132
$\texttt{NBJustWeb}_D$	**52.8%**	53.77	**52.8**	53.28	**264**	227	**9**

Table 4.25: Performance of the $\texttt{NBJustWeb}_D$ classifier run with the disambiguation procedure and comparison with $\texttt{BaseGR+Web}_D$, $\texttt{NB+Web}_D$, and $\texttt{NBStand}_D$ classifiers.

To verify this hypothesis, I tested all classifiers presented so far on a data set which contained NPs to the left of the correct antecedents as well. I used 408 samples of other-anaphors with antecedents in a two-sentence window and 10-fold cross validation. At each iteration of testing, 368 samples of other-anaphors were used for training and the remaining 40 cases for testing. Each sample was used in testing only once. In total, 400 samples were used in testing. I used different procedures to generate the training and test data sets. The training data was generated as earlier, i.e., the positive training instances were generated by paring each anaphor with its closest preceding antecedent. To generate negative training instances, I paired anaphors with each of the NPs that intervened between the anaphor and its antecedent in the two sentence window. (Thus there were no split antecedents in the training data set.) For testing, however, I kept all NPs in the two-sentence window, including those that occurred before the correct antecedent (and thus the test data set contained split antecedents). The total number of instances in the training data sets varied between 1,504 and 1,567 NPs. The test data sets contained each between 362 and 416 NPs, of which between 40 and 52 NPs were the correct antecedents. (The number of correct antecedents varied because of split antecedents.) All together, the test data sets contained 451 correct antecedents.

Table 4.26 presents the performance of the `Base1`, `NBBaseGR`, `NBBaseSEM`, `NBStand`, `NBStand+Web`, `NBBaseGR+Web`, and `NBJustWeb` classifiers on the testing data. (Precision, recall, and F-measure were averaged; the number of true positives and false positives were aggregated for all test data sets.)

As expected, the performance of the algorithms on the new test data set was signif-

Classifier	P	R	F	TP	FP
$\texttt{Base1}_{D00}$	27.75	27.75	27.75	111	289
$\texttt{NBBaseGR}_{400}$	31.62	32.33	31.52	146	326
$\texttt{NBBaseSEM}_{400}$	26.07	40.96	31.68	184	527
$\texttt{NBStand}_{400}$	26.19	47.03	35.56	211	525
$\texttt{NBStand+Web}_{400}$	30.84	**50.26**	38.07	**226**	510
$\texttt{NBBaseGR+Web}_{400}$	**39.86**	41.54	**40.32**	186	**286**
$\texttt{NBJustWeb}_{400}$	28.98	46.07	37.22	207	455

Table 4.26: Results for the `Base1`, `NBBaseGR`, `NBBaseSEM`, `NBStand`, `NBStand+Web`, `NBBaseGR+Web`, and `NBJustWeb` classifiers on a more realistic data set.

icantly lower than that on the biased data set. In particular, the precision of algorithms dropped dramatically. However, the results the classifiers achieved when run with the disambiguation procedure, were not as dramatically different from their performance on the biased data set. For instance, there was a significant difference in the precision of the $\texttt{NBBaseGR+Web}_D$ classifier on the old and new data sets (Table 4.27). The difference in the classifier's recall was, however, insignificant. Also, the trends were mostly the same for the new data sets as for the old one:

- Adding Web knowledge improved precision, recall, and F-measure of the `NBStand` classifier. (The difference was not significant.)

- Just grammar or WordNet semantics by themselves were not sufficient to resolve the antecedents of other-anaphors: the performance of the `NBBaseGR` and `NBBaseSEM` was lower than that of the `NBStand` classifier which used grammatical as well as semantic knowledge. (`NBStand` achieved significantly higher recall than `NBBaseGR`; all other measures, while higher for `NBStand` than for `NBBaseGR` and `NBBaseSEM`, were not significantly higher.)

- Grammatical features consistently showed higher precision than semantic features (both for the classifiers that use WordNet and the Web), but lower recall. Semantic features showed higher recall but lower precision than grammatical features.

- With respect to different semantic resources, the classifiers that used the Web as their source of semantic information (`NBBaseGR+Web` and `NBJustWeb`) achieved better results that the classifiers that used WordNet (`NBStand` and `NBBaseSEM`). The `NBBaseGR+Web` classifier showed significantly higher precision than `NBStand`; the recall was also higher, but not significantly higher. The `NBJustWeb` classifier outperformed the `NBBaseSEM` classifier. (The difference was not significant.) Altogether, this showed that using the Web as the source of semantic information leads to higher results than using WordNet. This was due to the following reasons: (1) the Web contained more information relevant for the resolution of other-anaphors, e.g., domain-specific and general world knowledge, which often was not available from WordNet; and (2) querying the Web with specific lexico-semantic patterns eliminated the need for prior word sense disambiguation and thus lead to more precise information.

- Grammar and semantics resolved different kinds of samples: `NBStand` outperformed `NBBaseGR` and `NBBaseSEM`, and `NBBaseGR+Web` outperformed `NBBaseGR` and `NBJustWeb` (in terms of F-measure; `NBJustWeb` actually achieved a higher recall than `NBBaseGR+Web`, but its precision was 10 percentage point lower.) This confirmed the hypothesis that, when resolving other-anaphors, it is necessary to consider several knowledge sources and that to get the best results, the sources should provide different types of information.

- The grammar had a "constraining" effect on semantics: the `NBBaseGR+Web` classifier produced fewer FPs than the `NBStand`, `NBStand+Web` or even `NBJustWeb` classifiers. (Although some of the false positives were due to WordNet and lack of word sense disambiguation, as the `NBBaseSEM` classifier produced as many FPs as `NBStand` and `NBStand+Web`.)

There was one difference in trends, however, for which I do not yet have an explanation: The `NBBaseGR+Web` classifier showed higher precision and F-measure than the `NBStand` classifier but lower recall. (This was not the case on a more constrained test data set). As the grammar component in the two algorithms was the same, the only difference was in their semantic features: Web vs. WordNet. And this is why

Classifier	Success rate	P	R	F	Correct	Incorrect	None
NBBaseGR+Web$_D$	**49.2%**	**66.85**	**49.2**	**56.68**	**246**	122	132
NBBaseGR+Web$_{D400}$	45.0%	59.8	45.0	51.35	180	**121**	**99**

Table 4.27: Performance of the NBBaseGR+Web$_D$ classifier when run with the disambiguation procedure on the old and new data sets.

this difference in performance is puzzling: NBJustWeb achieved higher recall than NBBaseSEM, but when combined with grammar, the recall was lower for the classifier that used Web knowledge than for the classifier that used WordNet-based knowledge.

4.6 Summary

In this chapter, I advocated that resolving antecedents of other-anaphors requires a probabilistic framework which considers several types of information at once: syntactic, semantic, recency, etc. To determine which framework would suit the phenomenon best, I compared the decision tree classifier commonly used in coreference resolution with the Naive Bayes classifier. I argued that the Naive Bayes classifier was more suitable for resolving the antecedents of other-anaphors because it consistently achieved higher recall while maintaining a satisfactory level of precision. Recall is as important as precision when resolving antecedents of other-anaphors, because, unlike, for instance, definite descriptions, other-anaphors are always anaphoric and, therefore, they require an antecedent. (More than half of definite descriptions are discourse-new.)

Having chosen Naive Bayes as the learning framework, I presented several classifiers that differed with respect to how many and what kind of knowledge they used. The NBStand classifier was, for instance, trained on standard grammatical and semantic features such as recency, syntactic role, and semantic knowledge from the WordNet lexical database. NBStand+Web, in addition to the standard features, used Web semantics and co-occurrence patterns. NBBaseGR+Web relied on grammatical information and Web knowledge to resolve antecedents of other-anaphors (i.e., no WordNet knowledge was used). I also considered a variety of baseline classifiers, from those using grammatical information only (NBBaseGR), to those relying primarily on semantic knowledge (NBBaseSEM and NBJustWeb), and a simple hand-crafted symbolic algorithm to indicate the difficulty of the task (Base1).

The major results reported in this chapter are as follows:

- Semantic knowledge (such as "steel is a commodity") is crucial in resolving other-anaphors. However, the quality and relevance of the semantic knowledge base are important to success. WordNet proved once again insufficient as a source of semantic information for resolving other-anaphora. Algorithms that used the Web as a knowledge base achieved better performance than those using WordNet, because the Web contains domain-specific and general world knowledge which is not available from WordNet.

- But semantic information by itself is not sufficient to resolve other-anaphors, as it seems to overgenerate, leading to many false positives.

- Grammatical features such as syntactic function of antecedent and anaphor, antecedent NP form, and distance have a "tempering" effect of semantics. The best results were obtained from a combination of semantic and grammatical resources.

I also pointed out an inherent limitation in the Naive Bayes model and briefly mentioned a way to handle it which I will explore in detail in the next chapter.

The error analysis of the Naive Bayes classifiers identified several issues that require further attention from the research community. These include the role of modification in anaphor resolution, the nature of bridging references and redescription samples and resources required to resolve them, the need for better NER, and challenges such as elliptic anaphors and samples with distractors. I also discussed the advantages and limitations of using the Web as source of semantic knowledge.

The performance of the best classifier, NBBaseGR+Web$_D$, was encouraging. However, the resolution procedure it used was not yet a full decision procedure. For instance, the classifier did not yet take into account the fact that other-anaphors always require an antecedent. Also, a full resolution procedure must take into account other factors, e.g., syntactic constraints on antecedent realization (Section 3.3.2). An approximation of such a full procedure is presented in the next chapter.

Chapter 5

A Hybrid approach to resolution of other-anaphora

One of the machine learning approaches to other-anaphora presented in the previous chapter achieved a good level of success on a rather difficult phenomenon. Specifically, the best performing classifier, $\texttt{NBBaseGR+Web}_D$, run with the disambiguation procedure, correctly resolved 49.2% anaphors, achieving an F-measure of 56.68. When tested on a more realistic data set (Section 4.5), the same algorithm, $\texttt{NBBaseGR+Web}_{D400}$, resolved correctly 45% of the antecedents. Its precision was 59.8, recall 45.9, and F-measure 51.32. While these results were satisfactory, it was clear that the classifier's recall could be raised somewhat and done so with relatively little effort. This chapter explores this issue. Also, the resolution procedure used by the classifier was not yet a full decision procedure. For instance, the classifier did not take into account the fact that other-anaphors always require an antecedent. Furthermore, the algorithm did not account for some absolute constraints on antecedent realization.

In this section, I present an approximation of such a full resolution procedure. The *hybrid* approach to other-anaphors described in this chapter combines a Naive Bayes learning model with a set of informed heuristics and back-off procedures (Section 5.1). The method is evaluated in Section 5.2, and in Section 5.3 the performance of the hybrid algorithm is compared with other approaches to other-anaphora.

5.1 The method

In Chapter 3.3.2, I identified four types of syntactic environments which cannot realize *both* other-anaphors and their antecedents at the same time. For instance, a noun phrase preceding an appositive cannot realize the antecedent of other-anaphor, if the appositive NP is the anaphor (Example 105). Likewise, an appositive NP cannot realize the antecedent if it modifies an other-anaphor (Example 106):

(105) Separately, a federal judge hearing Mr. Hunt's bankruptcy case yesterday turned down a proposed $65.7 million settlement between Mr. Hunt and Minpeco S.A., **another major creditor in the case**.

(106) **Another small Burgundy estate, Coche-Dury**, has just offered its 1987 Corton-Charlemagne for $155.

In Example 105, the noun phrase "another major creditor in the case" cannot be interpreted as "a major creditor in the case other than Minpeco S.A". (And most likely not as "a major creditor in the case other than Mr. Hunt".) Likewise, in Example 106, the anaphor "another small Burgundy estate" does not mean "a small Burgundy estate other than Coche-Dury".

Some factors that play a role in the interpretation and resolution of other-anaphors, e.g., distance between anaphor and antecedent, antecedent gender, NP form, or semantic class are *preferences*: in a particular data set, antecedents of "other" are more or less likely to be of a certain gender, NP form, and/or semantic class (see Section 4.3.5 for such preferences for the WSJ corpus). Also, these preferences might vary for different types of corpora, genres, and domains. Syntactic constraints, on the other hand, are *absolute* constraints. They apply to *all* antecedent candidates, across all genres, styles, corpora (of English; other languages might exhibit different types of syntactic constraints), and text domains. Therefore it is necessary to incorporate them into the decision procedure as a *filter*. In fact, an error analysis of the output of the $\texttt{NBBaseGR+Web}_D$ classifier showed that in at least half a dozen cases (and perhaps more) antecedents were resolved to incorrect entities which violated the syntactic constraints above. If the resolution procedure employed a syntactic filter, it is likely that the system would have made fewer incorrect predictions. The syntactic constraints have not yet been implemented; they were modeled through a manual inspection of the classifier's output.

To increase the number of true positives (and raise the classifier's recall), I used the WEB feature as a fall-back option. Recall from Section 4.4.2 that the `NBJustWeb`$_D$ classifier (trained on just the WEB feature) resolved — not necessarily correctly — more cases than e.g., `NBBaseGR+Web`$_D$. (It failed to predict *any* antecedent in just 9 cases.) Also, it correctly classified more antecedents than `NBBaseGR+Web`$_D$.

Another way to improve the classifier's recall was through resolution of examples which involved metonymic antecedents. Because of the nature of the articles in the WSJ corpus, the data set contains many samples of conventional metonymies such as `company-name-for-assets` (shares, stock, securities, certificates, etc.), e.g., Example 107.

(107) … *Mochida* fell 150 to 4,290. **Other losing issues** included Showa Shell, which fell 40 to 1,520.

Ultimately, conventional and other types of metonymies would require an expansion of the learning model, either through adding a special feature for metonymy, or though expanding the feature RELATION.[1] Instead, I employed a back-off metonymy resolution procedure, which was applied to samples in which the classifier had not predicted any antecedent. The procedure is as follows. If `NBBaseGR+Web`$_D$ returned no antecedent for a particular anaphor, the metonymy procedure scanned the text in a right-to-left fashion and proposed as antecedent the first NP which satisfied the following condition: the head noun of the anaphor was "shares", "stocks", "securities", "certificates", "bonds", or their synonyms, and the antecedent candidate was an organization.

By applying this simple heuristic to the samples, the algorithm gained four new correctly resolved cases. (Other cases with `company-name-for-shares` metonymy had been resolved prior to applying the metonymy resolution procedure.)

Altogether, the hybrid resolution procedure for other-anaphors is as follows (assuming that a classifier has been trained).

[1]The former alternative is more sound, as binary features result in stronger probabilities. Both alternatives would, however, require manual annotation. In the ML experiments reported in the previous chapter, all features values were acquired automatically.

Classifier	Success	P	R	*F*	Correct	Incorrect	None
$\texttt{NBBaseGR+Web}_D$	49.2%	**66.85**	49.2	56.68	246	**122**	132
$\texttt{Hybrid}_D$	**61.0%**	62.12	**61.0**	**61.56**	**305**	186	**9**

Table 5.1: Performance of the hybrid approach to other-anaphora on the original test data set and comparison with the performance of the $\texttt{NBBaseGR+Web}_D$ classifier.

1. Scan the sample from right to left and use the learned model to predict the antecedent of an other-anaphor, ignoring noun phrases which violate the syntactic constraints. Terminate the search when one antecedent has been found.

2. If the beginning of the sample has been reached and no antecedent has been found,

   (a) Resolve metonymies such as `company-name-for-assets`.

   (b) If the beginning of the sample has been reached and no antecedent has been found, choose as antecedent the first NP with the WEB value “webfirst”.

5.2 Results and analysis

Table 5.1 illustrates the performance of the hybrid resolution procedure on the original data set; the procedure’s performance on a more realistic test data set is given in Table 5.2. Because the two data sets differ with respect to how they were generated, it was necessary to modify the evaluation procedure for one of the sets. For the unbiased set only, the evaluation procedure was amended with the following condition. If the proposed antecedent was not the actual antecedent, but it formed a coreference chain with the actual antecedent, it was counted as correctly resolved. This was necessary because the data set contained cases with split antecedents (as a result of a larger window size).

Adding the metonymy and Web fall-back procedures significantly increased recall and success rate by 11.8 percentage points when tested on the original data set (23.98% improvement relative to the recall of $\texttt{NBBaseGR+Web}_D$). However, precision of the $\texttt{Hybrid}_D$ approach was 4.73 percentage point lower than that of the $\texttt{NBBaseGR+Web}_D$

Classifier	Success	P	R	F	Correct	Incorrect	None
NBBaseGR+Web$_{D400}$	45.0%	59.8	45.0	51.32	180	121	99
Hybrid$_{D400}$	**54.25%**	55.5	**54.25**	54.92	217	174	9

Table 5.2: Performance of the hybrid approach to other-anaphora on the unbiased test data set and comparison with the performance of the NBBaseGR+Web$_{D400}$ classifier.

classifier. (The difference was not significant.) Nevertheless, its F-measure was 4.88 percentage points higher (an 8.61% improvement relative to NBBaseGR+Web$_D$). On the unbiased test set, the results were as follows. Recall increased significantly by 9.25 percentage points (20.56% relative to recall of NBBaseGR+Web$_{D400}$). Precision dropped by 4.3 percentage points (7.75% relative to precision without the metonymy heuristic and Web fall-back procedure). (This difference was not significant.) F-measure increased by 3.6 percentage points (a 7.01% increase relative to F-measure of NBBaseGR+Web$_{D400}$).

Most of the gains were the result of the Web fall-back procedure, which gave 49 new correctly resolved cases on the original test data set and 37 new correct cases on the unbiased test data set. Below are some of these cases: "investors ... other holders", "U.S. government ... other groups", "1937-87 ... any other 50-year period since before the last Ice Age", "increasing costs as a result of greater financial exposure for members ... other , far-reaching repercussions", "this provision ... another measure", "home improvement items ... other big-ticket durable goods", "the model ... the other car", "designer's age ... other risk factors", "cross-connect systems ... other telecommunications equipment", "Sterling, Mass., grower ... another job", "an annual pension of more than $244,000 ... certain other fringe benefits", "strong performances in consumer durables and machinery orders ... other factors", "steel ... other commodities", "retiree shareholders and directors ... other workers", "colleges and universities ... other government units", "its proposed debt swap ... other alternatives for re-financing the debt", "bankruptcy reorganization plans ... other options for Eastern's future", and Example 108:

(108) This quarter's loss includes *pretax charges of $4.9 million on the proposed discontinuation of the company's troubled British subsidiary*, and $3.7 million of

other write-offs the company said were non-recurring and principally related to inventory, publishing advances and pre-publication costs.

Some of the samples that the hybrid resolution method resolved incorrectly were the familiar cases, listed below. Other errors were specific to the Web. As it is not always clear what caused an error — several factors may be at play — some of the groups below partially overlap with each other.

- Incorrect NE classifications lead to incorrect pattern instantiations, wrong co-occurrence frequencies, and ultimately to incorrect predictions:

  (109) *Wells Rich* declined to comment on the status of the account, as did **the other agencies**.

  (110) According to its most recent annual report ... Maxwell Communication bought $3.85 billion in assets — including *Macmillan Inc. and Official Airlines Guides* — and sold $2 billion in non-strategic businesses. Now, Maxwell founder Robert Maxwell says he has an appetite for new acquisitions in the U.S., adding that he could spend “a good deal more” than $1 billion on **another U.S. purchase**.

  In Example 109, the company “Wells Rich” was misclassified as a person. In Example 110, “Official Airlines Guides” was not recognized as a company name. This example is interesting for one more reason: some collocations are more “natural” than others; the phrase “acquisitions and other purchases” is very frequent, and therefore the antecedent was resolved to the NP “new acquisitions in the U.S.”.

- Example 111 is an example of redescription, and again as in Example 110, the phrase “measure and other indicators” is more frequent than “(employment) report and other indicators”.

  (111) *The employment report*, which provides the first official measure of the economy’s strength in October, is expected to show smaller gains in the generation of new jobs. **Other key economic indicators due this week** include ...

- Bridging cases are hard:

  (112) Erwin Tomash, the 67-year-old founder of this maker of data communications products and a former chairman and chief executive, resigned as a *director*. Dataproducts is fighting a hostile tender offer by DPC Acquisition Partners, a group led by New York-based Crescott Investments Associates. Under the circumstances, Dataproducts said, Mr. Tomash said he was unable to devote the time required because of **other commitments**.

  (113) Too often now, *a single court decision* becomes the precedent for **other, less compelling cases**.

  In Example 112, common knowledge suggests that being a director of a company is a (major) commitment. Still, collocations such as "time and other commitments" are a lot more common, while the pattern "director and other commitments" returned zero counts from the Web. Example 113, is hard to even paraphrase. In any case, what needs to be excluded from "other, less compelling cases" is not "a single court decision", but "a single case in which the decision was made". And again, as with many of the examples I quoted so far, the collocation "precedent(s) and other cases" is more frequent than "decision(s) and other cases".

- In Section 4.3.4.2, I showed that modification plays an important role in the interpretation of other-anaphors. Below are two more examples that illustrate this issue.

  (114) Earlier this year, Black & Decker put *three Emhart businesses on the auction block: the information and electronics segment, the Dynapert electrical assembly business and Mallory Capacitors*. *The three units* had combined 1988 sales of about $904 million. *The three units* contributed about a third of Emhart's total sales. In addition, Black & Decker had said it would sell **two other undisclosed Emhart operations** if it received the right price.

(115) Typically, he will be billed only *several weeks after the expenditure*, and then has **another couple of weeks** before he has to pay the bill.

In Example 114, when instantiating the Web search pattern, instantiating the anaphor with just “operations” returns — unsurprisingly — “sales and operations” as the highest scoring instantiation. In Example 115, the antecedent should be interpreted with respect to the PP complement of the anaphor, rather than with respect to its lexical head.

- Finally, hard cases such as Examples 89 and 99 in Section 4.3.4.3 are a challenge.

Other errors are specific to the Web approach:

- The hybrid algorithm failed to resolve only 9 cases of 500 and 400 respectively. All of these 9 cases contained pronominal antecedents, e.g.,

  (116) But board members say *he* took so long to decide how to vote that by the time *he* decided, it was too late to try to draw **other members** to his position.

  Such samples were unresolved because the Web look-up method used by Markert *et al.* (2003) (which I used to obtain MI counts to determine the WEB feature) automatically filtered out pronominal antecedents from the list of entities to look up on the Web and therefore they received the value “webrest”. Markert et al. filtered out pronominal antecedents because pronouns occur so frequently in texts that they would probably always get the highest counts, in particular because the scoring method that Markert et al. used takes into account individual frequencies of all terms in the search pattern. There is, however, no internal limitation in the Web lookup procedure with respect to pronouns. And a different scoring method, e.g., just using frequency scores without calculating mutual information, might circumvent this problem.

- Bugs in pattern instantiations, in particular in pluralizing the antecedents, lead to wrong predictions.

- And sometimes antecedent-anaphor collocations that intuitively made sense just did not happen to occur even in the largest corpus available to the community. In Example 117, for instance, the pattern “(bank) employees and other men” returned zero Web frequency.

  (117) Eight people ... were arrested in an investigation of an alleged drug money-laundering operation. The U.S. Attorney’s office filed a criminal complaint against *six bank employees* charging them with conspiracy in the scheme, which apparently was capable of handling millions of dollars a week by funneling cash through fictitious bank accounts. **Two other men** also were charged with participating in the operation .

5.3 Comparison with other approaches to other-anaphora

This section compares the hybrid approach to other-anaphora discussed in this chapter with three other methods. In Chapter 3, I presented two symbolic approaches to other-anaphora, LEX and SAL, that operate primarily on lexical information in WordNet and syntactic salience, respectively. The third approach was presented by Markert *et al.* (2003), who resolved antecedents of other-anaphors on the basis of frequency counts from the Web (Section 4.4). A comparison with other systems, e.g., for resolution of pronouns and definite descriptions, is not appropriate. Pronominal anaphora is governed by other constraints than other-anaphora. Definite descriptions are anaphoric in less than 50% of a cases, and thus require different resolution methods. In particular, systems that resolve definite NPs aim at high precision, while for other-anaphors, both recall and precision are of equal importance, since all other-anaphors always require an antecedent.

Table 5.3 compares the hybrid approach tested on the unbiased data set with the LEX and SAL algorithms, and with Markert *et al.* (2003)’s resolution method. The data sets used in the evaluation of the LEX, SAL, and Markert *et al.* (2003)’s algorithms are a subset of the unbiased data set on which the $\texttt{Hybrid}_{D400}$ approach was evaluated.

The hybrid approach outlined in this chapter achieved the best results so far on other-anaphora. Also, this is the first approach that combines machine learning and

Approach	Success rate	Precision	Recall	F-measure
LEX	49%	n/a	n/a	n/a
SAL	37%	n/a	n/a	n/a
Web-based	52.5%	n/a	n/a	n/a
$\texttt{Hybrid}_{D400}$	**54.25%**	55.5	54.25	54.92

Table 5.3: Comparison of the hybrid approach to other-anaphora with LEX, SAL, and the Web-Based algorithm of Markert *et al.* (2003). "n/a" stands for "not available".

symbolic methods in finding antecedents of other-anaphors. The approach requires a significant preprocessing effort in terms of acquisition of the necessary semantic and grammatical information. However, it is easy to integrate with an existing machine learning system for coreference, as most of the features it uses are similar to those used in coreference resolution (with the exception of the WEB and RELATION features).

The method would benefit from a better NER module and from further research into, e.g., the interaction between bridging and metonymic inferences; modification and its role in anaphora resolution; and improvements in the Web lookup method, which are subject for future work.

5.4 Summary

In this chapter I presented a hybrid approach to resolution of other-anaphora, which combined the Naive Bayes learning method described in Chapter 4 with a set of informed heuristics and back-off procedures. In particular, syntactic constraints on antecedent realization were incorporated in the decision process as a filter, rather than a preference, unlike, e.g., semantic class, gender and distance. Further, the decision process involved resolving one type of metonymy frequently occurring in the WSJ corpus, `company-name-for-assets`. (Metonymic antecedents are yet beyond what the learning component of the system can handle.) This heuristic was applied to cases in which the learner failed to predict an antecedent. Finally, I used the WEB feature as a fall-back mechanism: In cases in which the ML core of the system did not find an antecedent, the procedure used this feature as a fall-back, resolving the antecedent to the first NP with the value “webfirst”.

These heuristics significantly improved the method’s recall (while precision dropped insignificantly) and showed the strength of combining statistical and heuristics-based methods. These results for the hybrid approach to other-anaphora are the best results so far on this phenomenon.

Chapter 6

Conclusion

6.1 Summary and contributions

This thesis constitutes the first body of research into resolution of other-anaphora. Specifically, I focused on where and how to identify antecedents of other-anaphors, which is the first step towards their interpretation. I presented two symbolic, several machine learning, and one hybrid resolution approach to other-anaphora. The best performing approach was the hybrid approach that combined a probabilistic model based on the Naive Bayes classifier and a set of informed heuristics and back-off procedures. This approach evolved from a corpus study of other-anaphors in the British National Corpus, through the LEX and SAL symbolic algorithms, and a series of machine learning classifiers. This approach was compared with other approaches to other-anaphora (Section 5.3) and it achieved the best results to date on this phenomenon.

The approach presented in this dissertation focused on other-anaphors with NP antecedents, which is the most frequent antecedent type with other-anaphora (Chapter 2). While it is not uncommon to limit one's aspiration in this fashion, for instance, in the field of coreference resolution, most work to date has also been done on NP antecedents (Chapters 3 and 4), it would be desirable to have an approach that could handle all types of antecedents. More on this topic will be said below, in the section on future work.

Also, while quite successful, several of the methods presented in this dissertation were far from resolving all other-anaphors, and there is substantial room for improvement. For instance, the LEX and SAL approaches were not optimized to the same degree as the machine learning algorithms, not were they tested on a window size larger than two sentences.

And finally, the question that this dissertation attempted to answer is where and how one can find antecedents of other-anaphors. Identifying the correct antecedents of other-anaphors is only the first step towards their resolution. The second step involves interpreting the anaphors against their antecedents, and it requires working out the formal semantics of other-anaphora and the related issues of quantification, inference, bracketing ambiguities, and restrictive vs. non-restrictive modification which I touched upon in Section 2.7. All these issues are topics for future research.

In designing a successful anaphor resolution procedure it is necessary to know what factors play a role in determining antecedents of anaphoric expressions. This dissertation identified and evaluated such factors for other-anaphora. Specifically, I identified four types of syntactic environments which cannot simultaneously realize other-anaphors and their antecedents. These types of environments are absolute constraints on antecedent realization. There are also a number of factors which are gradient in nature: rather than allowing to completely exclude a NP from the set of antecedent candidates, they indicate a preference towards either interpreting it as antecedent or as non-antecedent. These factors and their relative contributions were examined through a series of experiments with symbolic and machine learning algorithms. Machine learning methods proved particularly useful for this task, as they did not require a commitment as to the order in which the features should be applied and because they allowed to treat the features as preferences.

Other contributions of this dissertation are less obvious than the feature analysis and resolution algorithms mentioned above. When I began working on other-anaphora, there were no corpora with other-anaphors and their antecedents annotated in them. I have since then produced two such corpora (and in the case of the BNC corpus, I also annotated a number of features that described the anaphors, antecedents, and relations between them). One of these corpora has already been used to test a competing resolution approach to other-anaphora (Markert *et al.*, 2003).

6.2 Future work

In addition to the issues mentioned above, there are five areas which I would like to explore in future work: (1) improving the machine learning framework; (2) extending the approach to anaphors with non-NP antecedents; (3) developing an analysis of bridging, metonymic, and redescription cases; (4) improving knowledge acquisition from the Web, and (5) testing the resolution procedure on corpora from other domains and languages and in a real information extraction or question answering systems. The subsequent sections contain ideas about how these topics might be pursued.

6.2.1 Improvements in the ML framework

There are several issues that require further attention with respect to the machine learning framework and feature acquisition. First, a better and more sensitive named entity recognition module is necessary. The NER module should not only be more effective in finding named entities and identifying their type but it should also cover a wider range of names than it currently does.

Second, the issue of NP modification needs to be addressed. I showed that some antecedents of other-anaphors are interpreted with respect to the anaphor's PP complement, rather than its head noun, and there are cases in which adjectival and nominal modifiers of other-anaphors provide important information which must be taken into account to get optimal resolution results. Likewise, restrictive relative clauses are essential in interpreting other-anaphors (while non-restrictive relative clauses are not), and it is necessary to develop a formal account of how they can be treated.

Third, binary features seem to result in stronger probabilities than features with many values. It would be interesting to see whether collapsing feature values into binary representations would lead to improved results.

Fourth, when examining posterior probabilities, I used 0.5 as the threshold of the likelihood for a particular NP to serve as antecedent of an other-anaphor. There were, however, several correct antecedents with the posterior probability just under 0.5. Lowering the probability threshold is an alternative strategy to the back-off procedure that the method currently uses.

Fifth and related to the issue above, to guarantee that the algorithm always selects exactly one antecedent, it would be interesting to test an alternative antecedent selection procedure. Rather than selecting as antecedents all NPs above a certain threshhold, as the algorithm currently does, it is possible to only select the NP with the highest posterior probablity. However, in addition to such a procedure, additional mechanisms must be put in place to handle cases with split antecedents.

Sixth, many samples of other-anaphors were resolved incorrectly even by the best performing algorithm because the texts contained distractors that occurred closer to the anaphor than the correct antecedents. Further research is necessary to identify methods to identify and exclude such entities from the set of potential antecedents. Alternatively, one might explore a method with different termination conditions. Currently, the search for an antecedent stops as soon as one antecedent has been found. (This antecedent may not be the correct one.) Besides terminating too early in some cases, the current search procedure is not capable of handling split antecedents.

Seventh, in addition to the 10 features I used, there might be other features that would be useful in identifying antecedents of other-anaphors.

Finally, I have argued that the Naive Bayes classifier is more suitable for resolving antecedents of other-anaphors than, e.g., decision trees, because it consistently achieved good precision without sacrificing recall. There are other ML methods which have been successfully used for a variety of NLP tasks and which might be suitable for resolving other-anaphora, e.g., Maximum entropy modeling, which has been used for coreference resolution with encouraging results (Kehler, 1997b). Maximum Entropy has an advantage over Naive Bayes, as it does not make any assumptions about the probability distribution and thus does not impose any additional constraints (e.g., feature independence) on the learned model. Another approach that seems worth experimenting with is the competition approach by Connolly *et al.* (1997), recently used by Iida *et al.* (2003) for resolution on Japanese zero pronouns.

6.2.2 Non-NP antecedents

All the algorithms in this dissertation were developed to handle other-anaphors with NP antecedents, which is the most frequent antecedent type. To be able to resolve all cases of other-anaphors in natural texts, it is necessary to extend the treatment to other types of antecedents, e.g., those realized as clauses, discourse segments, and various types of modifiers. Also, it is not unusual for the anchor of an other-anaphor to be mediated by the text or utterance situation without being explicitly mentioned. Such cases must be addressed as well. And finally, the resolution approach for other-anaphors should be integrated with a resolution mechanism for list-constructions and other-than constructions.

6.2.3 Bridging, metonymies, and redescriptions

Throughout the thesis, I have given many examples of other-anaphors with metonymic and bridging antecedents and examples in which the relation between the anaphor and antecedent is of redescription. All these examples have one thing in common — they involve a variety of inferential processes and therefore require a substantial amount of common sense, domain-dependent, and general world knowledge. Using the Web as a source of such knowledge has proved successful. There are, nevertheless, samples of other-NPs with which even the Web would not be much of help. A detailed analysis of these and other inferentially-heavy examples is needed, to understand what is involved in their interpretation and resolution. Parallel efforts are being undertaken to understand metaphoric and bridging inferences involved in the interpretation of definite descriptions, e.g., (Bunescu, 2003; Poesio, 2003) and the MASCARA project at the University of Edinburgh[1]. It is possible that the same inferential processes are involved in the interpretation of both types of anaphoric phenomena.

6.2.4 Knowledge acquisition from the Web

Semantic knowledge acquired from the Web turned out more relevant in resolving antecedents of other-anaphors than, e.g., knowledge available from WordNet lexical database. But Web knowledge also gave rise to incorrect resolutions primarily for two reasons. First, pattern instantiations submitted to the search engine could occur at a clause boundary, and because the returned results were not processed in any way, such

[1] http://www.ltg.ed.ac.uk/~malvi/mascara/

instantiations lead to incorrect associations (e.g., that bridges are people). This can be amended by refining the queries or by a shallow processing of the returned pages, e.g., by verifying that both the antecedent and anaphor entity in the pattern are of the same semantic class. Second, in some cases, it seems that antecedents of other-anaphors should be resolved to an entity with a relatively low mutual information score, rather than to an entity with the highest MI score. This may, however, not be the case. So far, my colleagues and I used only one construction — list-other — to acquire from the Web general- and domain-specific knowledge necessary to resolve other-anaphors. Other patterns, e.g., "X(s) such as Y(s)" and "X(s) other than Y(s)", can be used for the same purpose and, in fact, might boost the MI scores for some of the antecedents.

6.2.5 Interaction with real IE or QA system and testing on other domains and languages

The approach presented in this dissertation was developed and optimized on a relatively small corpus of samples from the *Wall Street Journal*. It is necessary to test it on a larger and more diverse data set, perhaps from a different domain or on general-purpose texts. Also, it would be interesting to see whether this approach could be ported to languages other than English, and how well it would integrate with a real reference resolution system, or information extraction or question answering system.

Bibliography

Aone, C. and Bennett, S. W. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 122–129, Cambridge, Mass, USA.

Asher, N. (1993). *Reference to Abstract Objects in English*. Kluwer Academics, Dordrecht.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 57–64, Las Palmas, Canary Islands.

Bierner, G. (2000). *Alternative Phrases: Theoretical Analysis and Practical Application*. Ph.D. thesis, The University of Edinburgh.

Bierner, G. (2001). Alternative phrases and natural language information retrieval. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, Toulouse, France.

Borthwick, A., Sterling, J., Agichtein, E., and Grisham, R. (1998). Exploting diverse knowledge sources via maximum entropy in named entity recognition. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 152–160.

Brennan, S. E., Friedman, M. W., and Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL'87)*, pages 155–162.

Bunescu, R. (2003). Associative anaphora resolution: A Web-based approach. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003) Workshop on The Computational Treatment of Anaphora*, Budapest, Hungary.

Byron, D. K. (2002). Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 80–87, Philadelphia, PA, USA.

Byron, D. K. and Allen, J. F. (1998). Resolving demonstrative anaphora in the TRAINS93 corpus. In *Proceedings of DAARC2 — Discourse, Anaphora and Reference Resolution Colloquium*, Lancaster University, UK.

Cardie, C. and Wagstaff, K. (1999). Noun phrase coreference as clustering. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 82–89, University of Maryland, MD, USA.

Carter, D. (1987). *Interpreting anaphors in natural language texts*. Ellis Horwood Limited, John Wiley and Sons, Chichester.

Chinchor, N. (1997). MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, Washington, DC, USA.

Cohen, W. W. (1995). Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning*, pages 115–123.

Connolly, D., Burger, J. D., and Day, D. S. (1997). A machine learning approach to anaphoric reference. In D. Jones and H. Somers, editors, *New Methods in Language Processing*, pages 133–144. UCL Press, London.

Corley, M., Corley, S., Crocker, M., and Keller, F. (1999). Gsearch corpus tools. http://www.hcrc.ed.ac.uk/gsearch/.

DeCristofaro, J., Strube, M., and McCoy, K. F. (1999). Building a tool for annotating reference in discourse. In *Proceedings of the ACL99 Workshop on the Relation of Discourse/Dialogue Structure and Reference*, pages 54–62.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.

Fraurud, K. (1992). *Processing Noun Phrases in Natural Discourse.* Ph.D. thesis, Stockholm University.

Ge, N., Hale, J., and Charniak, E. (1998). A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*.

Gordon, P. C., Grosz, B. J., and Gilliom, L. (1993). Pronouns, names and the centering of attention in discourse. *Cognitive Science*, **17**(3), 311–348.

Grosz, B. J. (1981). Focusing and description in natural language dialogues. In A. K. Joshi, B. L. Webber, and I. Sag, editors, *Elements of Discourse Understanding*, pages 85–105. Cambridge University Press, Cambridge.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1983). Providing a unified account of definite phrases in discourse. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1986). Towards a computational theory of discourse interpretation (manuscript).

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, **21**(2), 203–225.

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, **69**(2), 274–307.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

Harabagiu, S. (1998). *WordNet-based Inference of Textual Context, Cohesion and Coherence*. Ph.D. thesis, University of Southern California.

Harabagiu, S. and Maiorano, S. (1999). Knowledge-lean coreference resolution and its relation to textual cohesion and coherence. In *Proceedings of the ACL-99 Workshop on the Relation of Discourse/Dialogue Structure and Reference*, University of Maryland, USA.

Hawkins, J. A. (1978). *Definiteness and indefiniteness*. Croom Helm, London.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France.

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, **44**, 311–338.

Huddleston, R. and Pullum, G. K. (2002). *The Cambridge grammar of the English language*. Cambridge University Press, Cambridge.

Iida, R., Inui, K., Takamura, H., and Matsumono, Y. (2003). Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003) Workshop on The Computational Treatment of Anaphora*, Budapest, Hungary.

Joshi, A. K. and Kuhn, S. (1979). Centering logic: The role of entity centered sentence representation in natural language understanding. In *Proceedings of the International Join Conference on Artificial Intelligence*.

Joshi, A. K. and Weinstein, S. (1981). Control of inference: Role of some aspects of discourse structure — centering. In *Proceedings of the International Join Conference on Artificial Intelligence*, pages 385–387.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, Upper Saddle River, N.J.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic: Introduction to Model-Theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic, Dordrecht.

Kaplan, D. (1979). Dthat. In P. Cole, editor, *Syntax and Semantics 9, Pragmatics*, pages 221–243. Academis Press, New York.

Kehler, A. (1997a). Current theories of centering for pronoun interpretation: A critical evaluation. *Computational Linguistics*, **23**(3), 467–475.

Kehler, A. (1997b). Probabilistic coreference in information extraction. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language processing (EMNLP-97)*, pages 163–173, Providence, Rhode Island, USA.

Linde, C. (1979). Focus of attention and the choice of pronouns in discourse. In T. Givon, editor, *Syntax and Semantics 12, Discourse and Syntax*, pages 337–354. Academic Press, New York.

Markert, K. and Hahn, U. (2002). Understanding metonymies in discourse. *Artificial Intelligence*, **135**, 145–198.

Markert, K., Nissim, M., and Modjeska, N. N. (2003). Using the web for nominal anaphora resolution. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003) Workshop on The Computational Treatment of Anaphora*, Budapest, Hungary.

McCarthy, J. F. and Lehnert, W. G. (1995). Using decision trees for coreference resolution. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 1050–1055, Montreal, Canada.

Miller, G. A. (1998). Nouns in WordNet. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 24–45. The MIT Press.

Mitchell, T. M. (1997). *Machine Learning*. WCB/McGraw-Hill, Boston, Mass.

Modjeska, N. N. (2000). Towards a resolution of comparative anaphora: A corpus study of ”other”. In C. Broccias, M. Nissim, and A. Sanso, editors, *Semantics inside and outside the clause*. (in print).

Modjeska, N. N. (2002). Lexical and grammatical role constraints in resolving *other*-anaphora. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002)*, pages 129–134, Lisbon, Portugal.

Modjeska, N. N., Markert, K., and Nissim, M. (2003). Using the web in machine learning for other-anaphora resolution. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, Sapporo, Japan.

MUC-5 (1993). In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, San Fransisco, CA. Morgan Kaufmann.

MUC-6 (1995). In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, San Fransisco, CA. Morgan Kaufmann.

MUC-7 (1998). In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, San Fransisco, CA. Morgan Kaufmann.

Ng, V. and Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 104–111, Philadelphia, PA, USA.

Passonneau, R. J. (1993). Getting and keeping the center of attention. In M. Bates and R. M. Weischedel, editors, *Challenges in Natural Language Processing*, pages 179–227. Cambrige University Press.

Poesio, M. (2003). Associative descriptions and salience: A preliminary investigation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003) Workshop on The Computational Treatment of Anaphora*, Budapest, Hungary.

Poesio, M. and Modjeska, N. N. (2002). The THIS-NPs hypothesis: A corpus-based investigation. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002)*, pages 157–162, Lisbon, Portugal.

Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, **24**(2), 183–216.

Poesio, M., Ishikawa, T., Schulte im Walde, S., and Viera, R. (2002). Acquiring lexical knowledge for anaphora resolution. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 1220–1224, Providence, Rhode Island, USA.

Preiss, J. (2002). Anaphora resolution with word sense disambiguation. In *Proceedings of GLUK5*, pages 1–9.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole, editor, *Radical pragmatics*, pages 223–255. Academic Press, New York, NY.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman, London.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the 1st Conference on Empirical Methods in Natural Language processing (EMNLP-96)*, pages 133–142, Philadelphia, PA, USA.

Ratnaparkhi, A. (1998). Unsupervised statistical models for prepositional phrase attachment. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th Internationl Conference on Computational Linguistics*, pages 1079–1085, Montreal, Canada.

Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, **1**, 75–116.

Salmon-Alt, S. (2001). Interpreting other: From cognitive grammar to mutimodal dialogues. Paper submitted to SEMPRO-01: Workshop on Cognitively Plausible Models of Semantic Processing, University of Edinburgh, UK, 31 July 2001.

Sidner, C. L. (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. Ph.D. thesis, Artificial Intelligence Laboratory, Massachusetts Institute of Technology. Technical report 537.

Sidner, C. L. (1983). Focusing for interpretation of pronouns. *American Journal of Computational Linguistics*, **7**, 217–231.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, **27**(4), 521–544.

Staab, S. (1998). Grading knowledge — extracting degree information from texts. Ph.D. Thesis, University of Freiburg.

Staab, S. and Hahn, U. (1997). Comparatives in context. In *Proceedings of the 14th National Conference on Artificial Intelligence and 9th Innovative Applications of Artificial Intelligence Conference (AAAI'97)*, pages 616–621, Providence, Rhode Island, USA.

Strube, M. (1998). Never look back: An alternative to centering. In *Coling-ACL'98: Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the ACL*, pages 1251–1257.

Strube, M. (2002). NLP approaches to reference resolution. Tutorial notes, 40th Anniversary Meeting of the Association for Computational Linguistics, 7 July 2002, University of Pennsylvania, Philadelphia, PA, USA.

Strube, M. and Hahn, U. (1999). Functional centering — grounding referential coherence in information structure. *Computational Linguistics*, **25**(3), 309–344.

Strube, M., Rapp, S., and Christoph Müller, C. (2002). The influence of minimum edit distance on reference resolution. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 312–319, Philadelphia, PA, USA.

Tetreault, J. R. (2001). A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, **27**(4), 507–520.

van Deemter, K. and Kibble, R. (2000). On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, **26**(4).

Vieira, R. and Poesio, M. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, **26**(4), 539–593.

Walker, M. A. and Prince, E. F. (1996). A bilateral approach to givennes: A hearer-status algorithm and a centering algorithm. In T. Fretheim and J. K. Gundel, editors, *Reference and Referent Accessibility*, pages 291–306. John Benjamins, Amsterdam.

Webber, B. L. (1978). A formal approach to discourse anaphora. Bolt Beranek and Newman Inc., Report No. 3761.

Webber, B. L. (1991). Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, **6**(2), 106–135.