PRONOMINAL ANAPHORA RESOLUTION IN CHINESE

Susan P. Converse

A DISSERTATION

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2006

Dr. Martha Stone Palmer
Supervisor of Dissertation

Dr. Rajeev Alur
Graduate Group Chairperson

In loving memory

H. Peter Converse

Dr. Mary O. Amdur

Acknowledgements

First of all, my heartfelt gratitude to and deepest respect for the members of my committee: Aravind Joshi, Ellen Prince, Candy Sidner, and Mitch Marcus. Without their help this thesis would not exist.

They are the most recent of a long line of teachers to whom I am forever indebted: the Connecticut educators who decided to start teaching French to third graders; Mr. Nicholson, who took French beyond the classroom; Mr. Boynton, who first introduced me to the fun of linguistics and brought Latin alive; Mr. Merris and Mrs. Wilmot who taught us to write; Mrs. Walker who gave us the chance to do genetics experiments with *Drosophila melanogaster* and experience the enjoyment of discovering what the data had to tell us; Miss Yindrich who had us reading *Les Mouches* around the same time; Mr. Locke, who gave us beautiful and challenging music to sing and elicited our best efforts; Mrs. Schumacher and Miss Prescott, whose linguistics electives extended my horizons in linguistics; Mr. Homan who demanded critical thinking; Mrs. Maly, who encouraged my interest in math; Mr. Betterly, whose inspired teaching drew me into his Asian history elective, which led to wonderful new worlds, including Chinese; Helen T. Lin and Monica Yu, who taught me Chinese and more; Paul Cohen, who taught me the importance of a source's bias; Nathan Sivin, who nurtured my interest in early Chinese science; Dick Waters, who introduced me to Lisp and Emacs; Steve Cushing, whose NLP course started me on the path to graduate school; the professors here at Penn who have taught me so much during my studies here, Norm Badler, Robin Clark, Lila Gleitman,

Aravind Joshi, Mark Liberman, Mitch Marcus, Martha Palmer, Fernando Pereira, Ellen Prince, Mark Steedman, John Trueswell, Lyle Unger, and Bonnie Webber.

Thanks to mentor Charlie Muntz and especially to Mary Amdur, friend, teacher, and mentor.

My fellow students and the post-docs I have worked with have been my teachers as well, and I express my gratitude to them variously for their insights, technical help, native-speaker judgements, friendship, and moral support: Jan Allbeck, Olga Babko-Malaya, Gann Bierner, Ann Bies, Dan Bikel, John Blizer, Erwin Chan, Jinying Chen, David Chiang, Scott Cotton, Cassie Cresswell, Hoa Dang, Yuan Ding, Christy Doran, Miriam Eckert, Jason Eisner, Ryan Gabbard, Dan Gildea, Chung-hye Han, Na-Rae Han, Matt Huenerfauth, Elsi Kaiser, Alexandra Kinyon, Karin Kipper, Nobo Komagata, Seth Kulick, Jeremy LaCivita, Edward Loper, Xiaoyi Ma, Ryan McDonald, Eleni Miltsakaki, Tom Morton, Connie Parkes, Rashmi Prasad, Carlos Prolo, Joseph Rosenzweig, Anoop Sarkar, Andy Schein, William Schuler, Jonathan Shapiro, Libin Shen, Ben Snyder, Bangalore Srinivas, Debbie Steinig, Matthew Stone, Michael Strube, Fei Xia, and Szu-ting Yi.

For help with the laborious process of annotation, I am enormously grateful to Szu-ting Yi, Meiyu Chang, Jinying Chen, Kai-yun Chen, Ching-yi Chia, Fu-Dong Chiou, Tsan-Kuang Lee, Chia-ying Pan, and Nianwen Xue.

Thanks to my coaches, without whose help I never would have left grad school, gotten into grad school, and then successfully gotten out of grad school: Elissa Arons, Leie Carmody, Joe Wright, and Karen Groff.

To my neglected friends who have nevertheless supported me through these years I have profound thanks: Anja Stehr Carr, Nancy Cooper, Nancy Farwell and van and Saba, Susan Farwell and John and Abbott and Adrienne, Maggie Funderburg, Karen Larsen and Adli, Libby Levison, Thérèse Lung, Bill Needham, David Nicholas, and Adrienne Southgate and Jay and Scott.

My boundless love and gratitude to my local family Kirk and Yvonne and Kate

and Pete, who were always there for me.

For my long suffering family who have wondered if I would ever finish, thank you from the bottom of my heart for your patience with me and for your love and encouragement for so many years. Mom, yes, at last I am done.

Finally, there is no way I would ever have finished if it were not for the unending support and encouragement and help in innumerable ways from Mike Felker and Martha Palmer.

Thank you Mike for the countless rabbits you pulled out of a hat, and for your generosity of spirit and service. You always knew what had to be done and helped me do it. From the time I was first accepted at Penn you have smoothed the path. Your selflessness is a model to emulate.

Martha, without your guidance and encouragement I never would have made it. When I was discouraged about something you always had the right words to help me get over it and recharge and move forward. You came through for me with advice and support even when I picked the most inconvenient times to count on you. Your energy and enthusiasm and scholarship and warm heartedness have been invaluable to me. Unending thanks.

ABSTRACT

PRONOMINAL ANAPHORA RESOLUTION IN CHINESE

Susan P. Converse

Dr. Martha Stone Palmer

Resolving pronominal anaphors in English has been a focus of research in natural language processing for decades. Methods ranging from linguistics-oriented, rule-based approaches to data-oriented, machine-learning approaches have been applied to the problem of finding the antecedents of pronouns.

In contrast to the abundance of research in English, there is almost no work on the problem in Chinese. This thesis addresses that gap.

Both a rule-based and a machine-learning anaphora resolution approach are presented in this work. An important difference between Chinese and English is that Chinese, unlike English, is a pro-drop language, and has null (zero) pronouns. The rule-based approach is applied to resolving these null pronouns as well as to the overt, third-person pronouns.

The Hobbs algorithm is used for the rule-based method of anaphora resolution. Three versions of the algorithm are presented. The first uses only syntactic structure to select an antecedent. The second uses limited number and gender agreement, while the third incorporates semantic constraints on the proposed antecedents.

For the machine-learning method, maximum entropy, supervised machine-learning models are used. Different models were trained using sets of features that paralleled the information sources used by the different versions of the Hobbs algorithm.

Two sets of data were used. The Penn Chinese Treebank provided the test data for resolution of both overt, third-person pronouns and of zero pronouns. The CTB parses were annotated for coreference using guidelines that were drawn up for the work presented here. Data annotated for the 2004 Chinese ACE program were used for training and testing the maximum entropy models to find the antecedents for overt, third-person pronouns.

The results from experiments with the two basic methods using the different levels of linguistic information will be presented and discussed.

Contents

List of Tables

Chapter 1

Introduction

1.1 Background

People have an amazing ability to communicate with one another.

Human language is a complex, extremely powerful communication system. It does not merely provide the ability to signal, it enables its users to express an infinite number of new meanings.

Yet people take this ability for granted. We only appreciate the underlying complexity of it all when we watch children learning to speak, or when the system breaks down in some way. For example, we may mis-communicate with someone due to different assumptions, or we interact with someone who has suffered some language impairment from birth or accident or disease. We also appreciate the complexity of language when we try to communicate with someone who grew up in a culture that speaks a different language from our own.

A critical part of the machinery of language is the phenomenon of reference, that is, the process of using a linguistic symbol (written, spoken, or signed) to pick out a "thing" in the world, whether the "thing" is an object in the physical world like a glass of juice on a table or an abstract concept that only "exists" in our heads, such as the concept of "reference" itself. The linguistic symbol is called a "referring

expression." For the purposes of communication, the problem of reference is inextricable from the problem of establishing "common ground" – namely understanding what is mutually assumed and understood and known by two individuals who are trying to communicate with one another.

Why is the problem of reference important to a computational linguist?

If we want to make a computational system that can process human language in a way that is useful to people, we need to have a component that understands, not just one that correctly partitions the sound or text stream it is given as input into the correct language patterns (words, phrases, parses of sentences, etc.).

When we as human beings correctly understand something someone has said, it means we have successfully identified the entities the speaker has referred to as well as the relationships among them in a given context. We can then take actions based on that new knowledge. The action may be as simple as just storing a fact in memory, or as complex as making a series of inferences that result in fleeing a life-threatening situation. A computational system that understands at a minimum must recognize the entities in a discourse and establish the relationships among those entities in a context. "Recognizing" an entity means being able to link a referring expression in the text that refers to some real or abstract object with whatever internal representation the computational system is using for that object. The link may be straightforward or it may involve some computation, for example when inference or interpretation is required.

Not all referring expressions are created equal, however. They differ in their degrees of descriptiveness and specificity.

When we first introduce an entity into a discussion, we need to describe as much of it as we think is necessary for the hearer to understand which entity we are talking about. This may involve a lot of detail: "Do you remember the paperback book that I gave Aarika last year for her birthday that you thought was too young for her?" But once an entity is established in the conversation between two parties, often just

a pronoun will suffice to refer to it: "Yes, she thought it was silly." While names alone are usually sufficient without qualification on first mention, sometimes even a proper name will require expansion to pick out a unique individual: "Was that Ursula LeGuin or Ursula Andress who just won an award?"

The pair of sentences above about the gift book illustrate the problem of coreference, a crucial part of the more general problem of reference. Coreference occurs when two expressions refer to the same entity. Most of the time a hearer will have no problem understanding when an expression refers to an entity that was already mentioned.

> "Bennett got two CDs for his seventh birthday, *Peter and the Wolf* and *The Emperor's New Clothes*.
>
> He likes the first one better, but says the second is ok too."

We have no trouble understanding that "he" and "Bennett" corefer, that is, they both denote a boy in the world whose name is Bennett. Similarly, we figure out that "the first one" is coreferential with *"Peter and the Wolf"*, with both expressions denoting the actual CD of Prokofiev's music that the boy Bennett received.

The form of the referring expression is a very important part of our smooth comprehension.

This leads us to "anaphora resolution."

"Anaphor" comes from the Greek "to carry back." In computational linguistics, an anaphor is an expression that cannot be interpreted by itself. Instead, it depends for its meaning on finding the meaning of another expression in the discourse context. This second expression is the "antecedent" for the anaphor.

Thus comprehending an anaphor in a sentence requires figuring out which expression in the preceding context it depends on for its meaning, and then interpreting that expression.

Finding the antecedent expression for an anaphoric expression and interpreting the anaphor is called "anaphora resolution". Anaphora resolution and coreference

resolution are not the same thing, although they are usually conflated, and after this chapter will be conflated again, for convenience.

Coreference resolution is finding which referring expressions in a discourse refer to the same entity, even when the expressions are not anaphoric. For example, is "AT&T" the same entity as "American Telephone and Telegraph"? If so, the expressions are coreferential.

The antecedent expressions for anaphors may or may not be referring expressions.

Having tried to be clear about the distinction between these two processes, the following discussion will now return to muddy waters, and refer to a specific problem in which both coreference resolution and anaphora resolution overlap, pronoun understanding.

Pronouns are often anaphoric, in that they depend on an antecedent expression in the discourse context for their interpretation. Pronouns are usually also referential, in that the antecedent expressions on which they depend are frequently referring expressions that point to some entity. Thus talking about pronominal coreference resolution and pronominal anaphora resolution is equivalent in outcome most of the time.

Different forms of referring expressions, from pronouns to different types of noun phrases, have been categorized into an implicational hierarchy, called the GIVENNESS HIERARCHY, according to how salient their referents are in the discourse between speaker and hearer [17].

Givenness Hierarchy

in focus	>	activated	>	familiar	>	uniquely identifiable	>	referential	>	type identifiable
{*it*}		{*that*, *this*, *this* N}		{*that* N}		{*the* N}		{indefinite *this* N}		{*a* N}

The entity under discussion that is most salient under the right conditions can

be referred to using the most reduced expression "it", while an entity that is just being introduced must be represented by a form on the type identifiable end of the scale: "I saw *a car* at the car show that I really liked."

Psycholinguists have shown that people actually comprehend more quickly when the appropriate level of referring expression is used at each point in the discourse. Once an entity has been introduced into a discourse, it becomes discourse old, and there is no need, barring ambiguity, to keep on referring to it using expressions that are 'uniquely identifiable' or 'referential'. In fact, if a full noun phrase is used when a simple "it" will do, then the hearer may initially be misled into thinking that the entity is new, and take more time to realize that it is instead something familiar, already mentioned in the discourse (Clark and Wilkes-Gibbs, 1986 [6], Gordon *et al.*, 1993 [15]).

This means that a natural language understanding system is likely to encounter more referential expressions that are from the 'in focus' end of the scale than expressions that explicitly describe an entity, and it will have to "figure out" which abbreviated forms corefer with which definite descriptions that have come before it in the discourse.

It should be noted that, while the examples in the discussion thus far have been about concrete objects and people, many references are made that denote abstractions such as concepts (the notion of "reference"), propositions ("the sky is blue"), events (the spring equinox), or situations (the effects of a blackout of the power grid of the northeastern United States). These abstract "entities" may be part of "world knowledge" that most human beings acquire, they may be part of the shared world of the speaker and hearer, or they may be new propositions just introduced into the current discourse.

Computational linguistic researchers from the start have necessarily worked on the problem of coreference resolution in general, that is, establishing which referring

expressions of all types in a discourse refer to the same entity. Even though theoretical linguists have written about both coreference and anaphora for many languages, the bulk of the computational work until relatively recently has been in English, and much of that work has been on resolving coreferential relations between pronouns and nominal expressions, a limited subset of all referring expressions.

In the 1990's, computational linguists started publishing research on coreference resolution in Spanish and Japanese. What these two languages have in common that is different from English is that they are pro-drop languages, while standard English is not. That is, they allow a speaker to omit the subject in well formed prose sentences, given the correct context.

For example, Ferrández (1999 [12]) gives the following sentences.

Pedro$_i$	vio	a	Ana$_j$	en	el	parque.
name	SAW	obj	name	IN	THE	PARK

Pedro$_i$ saw Ana$_j$ in the park.

ϕ_j	Estaba	muy	guapa.
ϕ	WAS	VERY	BEAUTIFUL

She$_j$ was very beautiful.

The "dropped" subject in the second sentence is called by linguists a "zero pronoun" or a "null pronoun". In fact, just as an English sentence in context would sound more natural and would be easier to comprehend when a pronoun is used to denote a familiar/discourse old entity (note the use of "she" rather than "Ana" in the gloss), in Spanish or Japanese a zero pronoun subject would be more natural and easier to understand than if the sentence had an overt pronoun or noun phrase in place of the zero.

Chinese is another pro-drop language, with both overt pronouns and zero pronouns. Unlike Spanish or Japanese, however, it is not inflected. As in English, computational linguistic work in Chinese has been going on since there were computers but, perhaps due to the interesting problem of word segmentation in Chinese, which

added an extra problem to solve before the already challenging problem of parsing, most work to date has not yet focussed on coreference or discourse modeling.

This thesis addresses the problem of pronominal coreference resolution in Chinese. It will present methods for automatically identifying the coreferential relations not only for overt pronouns, but also for "zero pronouns."

1.2 Motivation – Why bother?

Why should we bother with pronominal coreference resolution, or even coreference resolution in general?

As mentioned above, once we introduce an entity into a discourse, if we continue to refer to that entity, we refer to it using shorthand expressions, not the full descriptive noun phrase or name used to introduce it, unless there are similar entities in the context that might cause ambiguity or we have pragmatic reasons for doing so.

Some common computational linguistic applications are systems that aim to answer questions whose answers may be found in a corpus or on the web, to summarize texts, to translate texts from one language to another, or to analyze text for interesting relationships among entities. Each of these systems needs not only to identify entities or topics of interest, but also to recognize when a new sentence presents a fact about an entity that has already been processed. Figuring out what expressions in a text refer to the same entity enables a system to correctly bind facts to the appropriate internal representations of the entities that have been recognized.

For example, a baseball fan or sports commentator might want to know "who was the 1990s Sporting News player of the decade?". The answer is found in the following passage, but can only be extracted if the pronoun "He" is correctly resolved to refer to the same referent as "Barry Bonds"[1].

[1]The quotation is from 2004 ACE nwire/NYT20001002.2133.0471.sgm

> Over a short postseason series, even the best hitters sometimes wallow through a slump as their team loses. Babe Ruth batted .118 (two for 17) in the five-game 1922 World Series that the Yankees lost to the New York Giants. But over 23 postseason games, the cream is supposed to rise to the top. Bonds' cream has only curdled. To his dismay.
>
> "I want people to say," he has often proclaimed, "that **Barry Bonds** is the best baseball player there is."
>
> If you just assess his career regular-season stats, you can argue that he is the best baseball player there is.
>
> **He** was named The Sporting News player of the decade for the 1990s over Ken Griffey Jr. and Mark McGwire, joining such prestigious names as Stan Musial ('40s), Ted Williams ('50s), Willie Mays ('60s), Pete Rose ('70s) and Mike Schmidt ('80s). In 10 years of MVP voting in the '90s, he received the most points – 1,895, nearly 500 more than the next highest vote-getter, Frank Thomas of the Chicago White Sox.

As another example, in translation from English to a language that inflects for gender, one would need to identify the antecedent of "it" in order to select the proper gender for the pronoun, or inflect the verb properly for a null subject. In the other direction, identifying the antecedent of a dropped subject could help in selecting among "he", "she", or "it" when translating to English from a pro-drop language such as Spanish or Chinese, as in this example from Hsin-Hsi Chen [3].

很	高	而且	,	腿	ϕ	很	修长	。	其中
Mary	very	tall	,	moreover	her	legs	very	slender	.

1.3 Statement of the Work

Broadly speaking there have been two kinds of approaches to pronominal coreference resolution, which here will be called "linguistics oriented" and "data oriented". The

linguistics-oriented approaches use rules and features based on linguistic theory to select the antecedent for an anaphoric expression, and tend to rely on a substantial amount of theoretically informed processing (including database construction) before the actual step of anaphora resolution can be performed. The data-oriented approaches use statistical or machine learning techniques that exploit the distributional reflections of linguistic phenomena to use in anaphora resolution.

The linguistics-oriented methods were the first to be used. They used hand-crafted rules that at a minimum were based on syntactic theory, such as the binding constraints (Chomsky, 1980 [5]), and gender and number agreement constraints. In addition to syntactic rules, some researchers ([10], [41], among others) added constraints from semantics and discourse structure to try to capture the correlation between the entity that was the focus of attention at a given point in the discourse and the form of expression used to denote that entity.

The linguistics-oriented methods have the chief limitation that because they are hand crafted, they are labor-intensive to build, and rule-based systems are not very robust when ported from one domain to another or from one genre to another, although much may be learned from a point of view of linguistic theory from moving to a new domain or genre and discovering what new phenomena must be addressed.

This shortcoming was one motivation for statistical approaches. When large annotated corpora became more widely available on-line in the 1980s, they provided enough data for statistical methods to be feasible. In addition, the availability of hand-parsed texts such as the Penn Treebank [32] provided linguistic gold-standard pre-processing "for free." Other linguistic resources such as Wordnet [11] provided useful extra-syntactic information.

Of course, while statistical or machine learning methods are flexible in that they can relatively easily be re-trained on corpora from new domains or genres, in another sense the cost in human effort has just moved from rule writing to data annotation. Instead of writing new rules for a new domain, effort is spent in hand parsing or

other annotation schemes.

From an engineering perspective, annotating corpora rather than writing rules to adapt a single system to a new domain or genre can sometimes be a productive allocation of effort toward solving natural language engineering problems in general. Once annotated, corpora may be used for any type of application using any type of model, linguistics- or data-oriented, and they can be used by many researchers. In contrast, writing more rules to make a single system more general or able to handle more domains is similar to extending an augmented finite automaton to parse a human language. Because people are infinitely creative in their use of language, the need for new rules will never end. A single system may achieve quite remarkable results, but if the system does not meet the needs of a broad class of users, it will have limited use.

This thesis explores both a linguistics-oriented approach and a data-oriented approach to the problem of pronominal coreference resolution, including zero pronouns, in Mandarin Chinese.

The "Hobbs Algorithm" (Hobbs, 1978 [20]) is used for the linguistics-oriented system. The algorithm has three incarnations in the work here. The first is a purely syntactic set of rules for stepping through parse trees. The second attempts to model number and gender constraints, which Hobbs actually included as part of the original algorithm, but which were treated separately here because Chinese has almost no number morphology, verbs are not inflected for gender or number, and gender is not marked. The third incarnation applies some semantic constraints to filter the answers proposed by the syntactic rules to achieve more accurate results[2]. This algorithm was designed for English, and has been used as a standard benchmark by many computational linguistic systems in English pronominal coreference resolution.

Why could it work for Chinese as well? It works for Chinese for both linguistic and engineering reasons. Linguistically, English and Chinese are both subject-verb-object

[2]and in his paper Hobbs also assumes a system that will do a lot more, such as handling split antecedents for plural expressions, ellipsis resolution, etc.

(SVO) languages. The core operational steps of the algorithm are top-down, left-to-right, breadth-first searches. The SVO nature of the languages matches this subject-biased search strategy. Chinese and English noun phrases differ in the locations of their heads, however. For example, the English

```
[ the book [that is on the desk]]
```

in Chinese would be

```
[[zhuozi shang de] shu ]
[[ desk   top  DE] book ]
```

with the English head "book" on the left, and the Chinese head "shu" on the right.

The fact that English noun heads are on the left and Chinese noun heads are on the right does not affect the choice the algorithm makes because it is working top down, and it stops at the topmost (outermost) NP. Both of the NPs above would meet the criteria, and the extraction of the head of each noun phrase is easily handled by language-specific support code.

The Hobbs algorithm is used to try to resolve both overt pronouns and zero pronouns in the Chinese Treebank (CTB) [53]), which has strings to denote empty categories, including dropped subjects (see the *Bracketing Guidelines* in [52]).

Maximum-entropy, supervised, machine learning models are used for the data-oriented approach. Because the amount of annotated data from the CTB was originally insufficient for statistical methods, data from the 2004 Penn Chinese ACE project are used for overt pronoun resolution (the automatic parses used in Chinese ACE did not produce empty categories, so zero pronouns are not addressed in these experiments).

Several experiments using the maximum entropy models were performed, using the same three levels of information that the three incarnations of the Hobbs algorithm use. The first experiments use feature sets that are primarily structural and

syntactic. Next, agreement features are added, and finally, semantic features are added to the training and execution of the models.

We hope to demonstrate the following with these experiments. First, that the simple, rule-based algorithm does reasonably well at resolving overt, third-person pronouns in Chinese. Second, that it can resolve matrix-level zero pronouns as well as overt pronouns. Third, that richer semantic features obtained automatically from existing linguistic resources can be used to approximate the third incarnation of the Hobbs algorithm to improve the performance. Fourth that, as in English, a maximum-entropy, machine learning model can achieve a good performance, and that using richer linguistic resources for features improves that performance.

1.4 Summary

This document is organized as follows.

Chapter 2 will summarize some of the previous research in coreference resolution methods.

Chapter 3 will discuss the annotation of the Chinese Treebank for pronominal coreference, and compare it to the MUC and ACE annotation schemes.

Chapter 4 will present the results of the experiments using the Hobbs algorithm for pronominal coreference when applied to the gold-standard CTB parses for both third-person pronouns and the null pronoun, and when applied to the overt, third-person pronouns in the machine parses of the 2004 Chinese ACE texts.

Chapter 5 will discuss the maximum entropy models that were trained and run in the Chinese ACE environment for resolving overt, third-person pronouns.

Chapter 6 will summarize the results and discuss future work.

Chapter 2

Previous Work

We now live in a resource-rich computational environment. We have huge amounts of memory and storage space available to us, and we have clusters of processors we can harness to work in parallel on big problems. On top of this, we have a wealth of texts on-line for NLP, there are new language resources made available by the day, and the World Wide Web provides an infinite source of material for almost anything that one could think of.

This was not always the case for natural language processing (NLP) researchers. In the early years, processing millions of words of text would be an enormous effort, even if the million words were actually on mag tapes someplace, which they weren't.

There always has been a tension within the realm of artificial intelligence (AI) between the two goals of trying to discover models that emulate how human beings do what they do and trying to create programs that achieve what human beings achieve, but not necessarily in the same way that humans do.

Now, with the availability of "unbounded" data and computing resources, we have a second tension that sometimes appears to be pulling us in two directions (and that computational linguists of the 1970s (or 50s) could only have dreamed about). That choice is whether to approach a problem from the point of view of using and pushing the boundaries of what linguistic theory and cognitive science say

about how human language processing works or might work, or to approach it from the data side, that is, using statistical processing on the phenomena we observe to try to develop or learn solutions to problems.

As in the "classical" AI tradeoff, the two approaches are not mutually exclusive, and one might venture to wager that the most successful researchers working from one perspective are those who are tuned in to the most recent advances from the researchers working from the other point of view. Our brains are amazily good at finding patterns, and for all we know a structural rule that is generally accepted in linguistics theory (for example one of the binding constraints in GB theory) might very well have a statistical implementation in the brain.

In Chapter 1 these two approaches were referred to by the labels "linguistics-oriented" and "data-oriented".

The research in anaphora resolution in general, and pronominal anaphora resolution in particular, has followed the trend in AI and NLP research from rule-based systems to statistical methods. But the trend is not absolute or one way. As a former professor said, "the world is not black and white, it is grey."

2.1 Early work

The early work in anaphora resolution, usually approached as part of solving a larger problem, used heuristics or was rule-based in some way[1].

In her 1983 chapter [41], Sidner categorized previous work by the kinds of linguistic knowledge brought to bear on solving the problem: heuristics, syntactic and semantic constraints, inference, and discourse context.

What the works she cited had in common was a large effort in hand coding knowledge, both procedural in the form of linguistic constraints and inference mechanisms, and static in the form of real-world facts.

[1] In which *rule-based* is taken to mean there is some theory or principle behind the rules, while heuristics are more *ad hoc*, i.e., whatever worked for the particular situation).

2.2 Middle traditional work

2.2.1 Constraints and Preferences

The anaphora resolution strategy of first constraining the set of possible antecedents by throwing out the "impossible" cases (e.g., using syntactic rules such as number and gender agreement, or semantic constraints such as the fact that castles don't move), and then choosing from among the set of remaining candidates by some preference mechanism, was used in a variety of ways by several rule-based systems.

The most comprehensive of these was the RAP system of Lappin and Leass [30], who set a high standard with their comprehensive approach.

2.2.2 The Centering Theory Thread

The original papers (published and circulated) on Centering Theory, beginning in 1983[2], did not propose the theory as a model for pronominal anaphora resolution *per se*. Rather the theory was aimed at investigating the relationships among "focus of attention, choice of referring expression, and perceived coherence of utterances within a discourse segment." [16].

Nevertheless, in her 1985 thesis, Kameyama combined the Lexical Functional Grammar (LFG) formalism with Centering Theory to outline a "zero pronominal interpretation model" for Japanese. The framework she proposed assumed both a source of commonsense knowledge and an inference component.

But because the notions of the coherence of a text/dialogue, the choice of the form of referring expression, and the focus of attention of the hearer (and speaker) are all inseparable, one of the rules in the theory was Rule 1, dubbed the "Pronoun Rule".

The Pronoun Rule effectively said that if a pronoun is used in an utterance to refer to an entity, then any entity in that utterance that is more salient given the

[2]According to the introduction and footnote 1 of (Grosz, Joshi, and Weinstein, 1995 [16])

context from the previous utterance must also be referred to using a pronoun. It was inevitable that for the pronominal form of referring expressions, therefore, that someone would think of working the theory "backwards" and try to use it in a discourse-motivated algorithm for resolving pronominal anaphors.

In 1987 Brennan, Walker-Friedman, and Pollack proposed an algorithm (dubbed the BFP algorithm in the literature) "to track discourse context and bind pronouns." In the process, they also proposed an extension to Centering Theory to handle certain sequences of utterances that had multiple ambiguous pronouns.

The extension to the theory was to split the original theory's single "shift" transition into "shift-1" and "shift" (later renamed "smooth shift" and "rough shift", respectively, by Walker, Iida, and Cote (1994)[47][3]).

In addition to using the constraints and rules of the Centering Theory to order potential anaphor-antecedent pairs, rule out certain candidates as antecedents, and rank the remaining candidates, they also incorporated a syntactic contraindexing filter to potential antecedents before the ranking step.

Brennan *et al.* did not give results for the performance of their algorithm in their paper, but Walker (1989 [48]) did a hand simulation to evaluate the algorithm in comparison with the Hobbs algorithm[4] on the same texts.

She found that the Hobbs algorithm outperformed BFP over domains consisting of newspaper articles (89% to 79%) and in a task domain 51 percent compared to 49 percent.

In a short article in 1997, Andrew Kehler [28] pointed out a limitation in the algorithm with respect to the garden path effect. The underlying problem is that, because the interpretation of a pronoun in the algorithm depends both on identifying the backward looking center, and on the preferred ordering of transition states from Rule 2 of the theory, it is not possible to resolve a pronoun before processing the entire utterance containing it. Thus the algorithm

[3] as cited by Kehler, 1997 [28].

[4] as presented in Hobbs, 1976 [19], an earlier version of the 1978 paper [20]

> "cannot model an addressee's immediate tendency to interpret a pronoun, and therefore cannot properly account for the pronoun interpretation preferences that result from such tendencies."

In 2001 Joel R. Tetreault evaluated his own Left-Right Centering algorithm (Tetreault, 1999 [45]) in comparison with the BFP algorithm, the Hobbs algorithm, and Strube's S-list approach [44]).

He implemented all four algorithms and executed them on 1694 unquoted pronouns appearing in 195 articles from the *New York Times* and 511 unquoted pronouns taken from fictional texts in the Penn Treebank.

Table 2.1:

	NYT		
	success rate	**% right intra-**	**% right inter-**
BFP	59.4	75.1	48.0
S-list	71.7	74.1	67.5
LRC	74.9	72.0	82.0
Hobbs	76.8	74.2	82.0
LRC-F	80.4	77.7	87.3
	Fictional Texts		
	success rate	**% right intra-**	**% right inter-**
BFP	46.4	18.8	43.8
S-list	66.1	84.4	56.5
LRC	72.1	84.3	64.2
Hobbs	80.1	85.8	75.2
LRC-F	81.1	86.0	76.2

For success rate, he used the number of successfully resolved pronouns divided by the total number of pronouns.

The original **LRC** ranks the Cf-list by grammatical function, while the **LRC-F** starts by ranking the Cf-list by grammatical function, but moves entities in a prepended phrase to the back of the Cf-list.

He ran the Hobbs algorithm with gender and number agreement but without selectional constraints

Centering in Chinese

While there has not yet been any work that used an implementation of the BFP algorithm or one of its children or alternates on Chinese, there has been linguistic work on Centering in Chinese. Most notably, Megumi Kameyama, in her seminal thesis on Zero anaphora in Japanese ([27]), discussed the discourse functions of zero pronouns and overt pronouns in Chinese in the course of her exploration of their roles in Japanese.

Recently, Zhiyi Song explored subject drop in Old Chinese *vs.* Modern Chinese with respect to Centering, and found that Continue and Smooth-shift Transitions of the theory favored null subjects, while Rough-shift and Retain Transitions *dis*favored null subjects.

2.2.3 Do we really need to parse the *whole* sentence?

The methods mentioned above rely on a lot of processing and a lot of hand crafting of rules and data, so researchers started to push the boundaries of what they could get away with *not* doing. This strategy was dubbed the "knowledge poor" approach.

Kennedy and Boguraev [29] tried a parserless implementation of Lappin and Leass's RAP algorithm.

Breck Baldwin [2] also implemented a shallow parsing approach.

2.3 Data-oriented approaches

As researchers who were following the linguistics-oriented approaches were exploring the lower bounds of the necessary linguistic processing, others were looking at ways

to use the distributional frequencies of different linguistic phenomena in order to automate tasks that were time consuming to do by hand.

As will be discussed in more detail in Chapter 4, the Hobbs algorithm showed better performance when semantic selectional restrictions were used to filter the antecedents that it proposed. But a database of semantic features with broad coverage is not easy to come by.

Dagan and Itai

Dagan and Itai [9] proposed using collocation information gathered automatically from a large corpus in lieu of a semantic database for filtering antecedents in the Hobbs algorithm.

They tested their method on 74 instances of the pronoun "it" alone, in order to have a pronominal anaphor with the most potential antecedents and the fewest a priori constraints, and limited the instances to cases in which "it" was coreferential with a noun phrase.

The idea was that if the proposed antecedent nominal appeared frequently in the same kind of predicate-argument context as the anaphor, then it was a more likely antecedent than if it did not.

To collect the counts for cooccurrence patterns they used a slot grammar parser that provided predicate-argument structures. The patterns used the actual literal words, not semantic classes for words.

Dagan and Itai modified the Hobbs algorithm to not stop after it found an antecedent to propose, but to keep running to propose two more antecedents (the choice of three was arbitrary). The three candidates were kept in the order in which they were "found".

For the pattern filter to apply, they required that the pattern count be three or greater for at least one of the three antecedents proposed by the algorithm. They found that the cutoff of at least three was useful to ignore noisy data.

The second or third proposed antecedent would be preferred if one of its corresponding patterns occurred at least twice as often as the first candidate, or if the first candidate did not have a pattern over the threshold.

Sentences containing the word "it" were randomly selected from a collection of articles from the *Washington Post*, and the immediately preceding sentences were extracted as well.

The following cases were then thrown out: sentences of over 25 words; bad parses; cases that did not have nominal antecedents with common noun heads; non-referential "it"; cases in which there was only one syntactically possible referent. This left 74 cases of ambiguous references to non-anaphoric common nouns in sentences that had been sucessfully parsed.

Patterns were collected from 40 million words of *Washington Post* articles, 24 million words of Associated Press news wire, and 85 million words from the Hansard corpus. Statistics were only collected from sentences of 25 words or less that had successful parses.

There were only In only 36 of the 74 cases that had patterns with counts over the threshold count of three. When Hobbs was run without using the statistical filter, it found the correct antecedent for 23 of the pronouns, for a 63.9% accuracy. When the pattern filter was applied, 31 of the proposed antecedents were correct, or 86.1% accuracy.

Although Dagan and Itai did not give the counts for the 38 cases that did not meet the threshold, they said that the improvement from the pattern filter for all 74 cases was from 64% without it to 74% with the filter.

All data-driven

Other efforts relied on statistical methods for the entire effort.

Aone and Bennett [1] used decision trees in a machine-learning application.

Ge, Hale, and Charniak [14] used the probabilities of a combination of linguistic

features to pick the most likely antecedent.

Tom Morton [36] demonstrated the contribution that coreference resolution made to improving question answering.

2.4 Does everyone in the world just speak English?

In the 1990s there was a flourishing of work on anaphora resolution in languages other than English. Mitkov (2002) [34] cites work in French, German, Japanese, Portuguese, Spanish, and Turkish.

To pick two pro-drop languages in different family trees we will just mention some work in Spanish and Japanese. Notably, unlike Chinese, both languages have rich inflectional systems.

2.4.1 Zeroes in Spanish

In the context of a machine translation system implemented in Prolog, Ferrández, Palomar, and Moreno (1999 [12]) outlined a pronoun resolution algorithm that they applied to both overt third person pronouns and demonstratives and to zero subjects that they detected automatically.

Their approach was one using constraints and preferences.

They used a slot unification grammar parser that would produce either full parses or partial parses according to a runtime parameter. For this paper they used partial parses. The parser produced "slot structures" that had empty slots for unfilled arguments. These were used to detect zero pronouns for verbs that were not imperatives or "impersonal" (e.g., "Llueve."/"[It] rains.").

They used an algorithm that applied a series of constraints to prune the list of potential antecedent candidates, then a few preferences to sort the remaining

candidates into a list, and then unified the anaphor with the first NP on the list. Constraints included gender, number, and person matching, followed by c-command and binding constraints (for the partial parses they used heuristics in lieu of the actual c-command relations)[5].

Preferences included the number of times the candidate had been mentioned in the preceding text, as well as surface structure phenomena such as parallelism – although the degree to which these were used in the experiments was not stated.

The corpus used was the Spanish version of *The Blue Book* corpus which contains the International Telecommunications Union CCITT handbook. They tested the algorithm on 100 Spanish third-person pronouns and demonstratives, and obtained an overall accuracy of 83 percent.

The breakdown of the syntactic placement of the pronouns and the accuracy of the algorithm for that group of 100 is as follows

Table 2.2:

Type of pronoun	**Percent correctly resolved**
53 were complement pronouns (direct objects)	85%
26 pronouns that were inside PPs	85%
21 pronouns that were not inside PPs	76%

When they looked at what contribution the constraints made to their results, they discovered that the morphological and c-command filters together reduced the total of 2,210 antecedent candidates proposed by the system down to just 906, a 59% reduction.

The c-command filter alone only reduced the number to 2,062 (and the algorithm run on those candidates only achieved an accuracy of 58 percent). The person-number-gender filter reduced the number of candidates to 993, and the algorithm was correct 73% of the time when those candidates were all considered.

[5]The paper discusses a semantic ontology linked to their lexicon, but they did not use it in producing the results for the paper.

For testing the zero pronouns, they used the Lexesp corpus, which contains Spanish texts from different genres. It has 99 sentences, containing 2,213 words, with an average of 21 words per sentence.

Their heuristics detected 181 verbs, of which 75% had a missing subject, and their system resolved 97% of those subjects correctly.

A later paper building on the same system [40] they reported a success rate of 76.8% in identifying antecedents for 1,677 third person personal pronouns, demonstrative pronouns, reflexive pronouns, and zero pronouns, taken from the same Blue Book and Lexesp corpora.

2.4.2 Japanese

There has been a lot of computational work anaphors in Japanese of all kinds, including zero pronouns. Just a sampling includes works such as the following.

Mori *et al.* [35] did work on resolving zero pronouns in instruction manuals exploiting the structure of the manuals as well as constraints based on syntactic patterns, including verb forms and types of discourse connectives. They reported 80.6% precision in identifying zero subjects.

Naikawa [39, 38] did work on finding rules for anaphora resolution of Japanese zero pronouns by using aligned Japanese-English sentence pairs.

Murata [37] used rules based on example, surface structures with key lexical items or forms, to resolve demonstrative pronouns, personal pronouns, and zero pronouns. They reported a precision rate of 78% for their test sentences.

In their 1990 paper, Walker, Iida, and Cote [49] discuss the interpretation of zero pronouns in Japanese based on Centering Theory. The paper is really focussed exploring how Centering Theory works in Japanese rather than on developing a anaphora resolution algorithm. A follow-up paper in 1994 [47] expands greatly on the 1990 work.

2.5 Chinese

While there has been an abundance of theoretical linguistic work on anaphora, and zero anaphora, in Chinese ([21, 22, 23, 24, 25, 26, 42])[6], computational work in this area has only recently begun.

Anaphors in translation

In 1992 Hsin-Hsi Chen did an analysis of anaphors in the context of transfer-based machine translation, both Chinese-to-English and English-to-Chinese (using Government and Binding theory for the syntactic formalisms). His definition of "anaphor" was loose, including reflexive pronouns, overt pronouns, big PRO, small pro, traces, and "variable" anaphors.

The Chinese reflexive 自己 is genderless and numberless, but in English reflexives agree in gender and number with their antecedents. Thus for the Chinese-to-English case, resolution is necessary for correct translation. Chen proposed a simple Reflexive Resolution Algorithm based on Binding Principle A and a "subject-orientation condition" that tried to find an NP in a subject position "as top level as possible". His algorithm also required a check that the antecedent be an animate noun.

He did not detail in his paper where the semantic information would come from, nor did he provide any evaluation of the performance of any of the algorithms he gave in the paper.

Chen used the subcategorization frames of the verbs to determine where the empty categories would appear, but did not give any rules for determining the antecedents of those that were truly anaphoric.

What he did do that was useful was to annotate a small parallel corpus and give a distribution of the zero and overt pronouns.

The corpus was selected from the Bilingual Sections of *Reader's Digest* in the Chinese version. He annotated each noun phrase in the texts with a tag of the form:

[6] Huang is to Johnson as Chinese is to English

```
(index1, type, environment, syntactic role, semantic case, index2)
```

In which he assigned the following kinds of values to the tag items:

- **index1** was the index of the noun phrase, and **index2** was the index of its antecedent (or '_' if it was not an anaphor)
- **type**: PRO, zero, pronoun, proper noun, definite NP, indefinite NP, bare NP, demonstrative NP, etc.
- **environment**: coordination, subordination, predicate construction, verb category, etc. (i.e., discourse context)
- **syntactic role**: subject, object, prepositional object, etc. (i.e., syntactic function)
- **semantic case**: agent, patient, source goal, etc. (i.e., thematic role)

Note that the data below barely use this coding, and use extra information that is not in the coding, unless there is more to 'environment' than he mentioned.

He presented the results of the distances between an anaphor and its antecedent for English and Chinese for four of the texts in terms of where the antecedent appeared relative to the anaphor.

In English both zeroes and personal pronouns preferred antecedents that were nearby: zero antecedents were preferred within the same sentence (which is a good thing since it is ellipsis) and personal pronouns in the previous sentence, but a next best thing was the previous clause.

Most possessives in English appeared in the same clause.

Chen proposed that the reason that zero anaphors in Chinese outnumbered the zero anaphors in English was that "pronouns are not repeatable in Chinese, while they are in English. Instead of repeating overt pronouns, zero pronouns should be used."

distance –> type ↓	same clause	previous clause	same sentence	previous sentence	same paragraph	previous paragraph	Total
English							
zero anaphor	1 4.4%	22 95.6%					23
personal pronoun	6 7.9%	23 30.3%	7 9.2%	30 39.5%	3 3.9%	7 9.2%	76
possessive pronoun	26 56.5%	8 17.4%	3 6.5%	7 15.2%	1 2.2%	1 2.2%	
Chinese							
zero anaphor		80 72.7%	14 12.7%	13 11.8%	2 1.8%	1 0.9%	110
personal pronoun	5 7.8%	19 29.7%	7 10.9%	26 40.6%	1 1.6%	6 9.4%	64
possessive pronoun	3 23.1%	2 15.4%	3 23.1%	4 30.8%		1 7.7%	13

The distribution of personal pronouns is similar in Chinese and English, but Chinese has fewer instances of possessive pronouns for the same texts. Chen says that one possible reason for this is that Chinese uses a lot of bare NPs in place of possessive NPs.

From his analysis of the syntactic functions of the antecedents of the zero anaphors and personal pronouns, he found that subject position was most preferred in both Chinese and English, and objects were preferred over objects of prepositions.

Table 2.3: **The syntactic roles of antecedents in English and Chinese**

Type	**subject**	**object**	**prep. object**
English			
ZA	21 / 91.3%	2 / 8.7%	–
PA	80 / 65.6%	24 / 19.7%	18 / 14.7%
Chinese			
ZA	90 / 81.8%	15 / 13.6%	5 / 4.6%
PA	57 / 74.0%	12 / 15.6%	8 / 10.4%

When he analyzed the "environments", namely the discourse contexts, of the anaphor-antecedent pairs for zero anaphors, he found that 57.3% (63 of 110) were subject-subject pairs appearing in coordinated discourse contexts.

Zero pronoun detection and resolution

More recently, Ching-Long Yeh, who did work on generation of zero anaphora in Chinese has done some work on detecting and resolving zero pronouns in Chinese [54].

He did not use complete parses, but instead did part-of-speech tagging followed by phrase-level chunking. From the chunked sentence he then created data structures he called *triples* to be used both in detecting zero pronouns and in resolving them.

Yeh tested on a corpus of 150 news articles containing 4,631 utterances and 41K words.

He reported an 80.5% precision in detecting zero pronouns, and a recall for zero pronoun resolution of 70% with a precision for resolution of 60.3 percent.

Of the total zero anaphors, therefore, a very small number is being correctly resolved: 0.80 x 0.7 x 0.6: roughly 34 percent.

Even with 100% detection of zero pronouns, the number of zeroes correctly resolved would be only 42%. There clearly is room for improvement in these numbers.

Chapter 3

Annotation Guidelines

3.1 Motivation and Scope of the Annotation

To evaluate any automatic method for resolving pronomonial anaphors it is necessary to have a set of test cases with their answer keys so that the output of the method may be scored. In addition, to develop supervised machine-learning methods of anaphora resolution requires training examples that are annotated with the answers.

A corpus that is annotated to indicate the relation between anaphors and their meaning-bearing antecedents is therefore in general desirable, and the larger the corpus the better for use with statistical methods.

Because one goal of this research is to address the problem of automatically resolving zero-pronoun anaphora, and the parses of the Penn Chinese Treebank (CTB) have specific strings to denote empty categories, including dropped subjects, the CTB was chosen as the data set for annotation. The string used in the CTB for the zero pronoun is "*pro*", and ***pro*** will be used throughout this document to refer to zero pronouns.

The anaphor-antecedent relations are annotated by coindexing the two syntactic expressions.

This work is limited to finding antecedent expressions for those anaphors that

refer to nominal entities, but not all pronouns (zero or overt) in Chinese necessarily have antecedents that represent nominal entities.

For example, just as English has pleonastic "it" ("It's raining.") or existential "there" in constructions such as "There are five books on the table.", in Chinese a zero pronoun will be used with the verb 有/"to have":

桌子/table 上/on ***pro*** 有/have 五/five 本/M 书/book 。/.

Another common situation in which a pronoun does not have a nominal antecedent is when the antecedent of the anaphor is a clause expressing an entire proposition or a phrase describing an event. These discourse anaphoric pronouns are labeled "discourse deictic" [50] and will be discussed below in Section 3.7.1.

In order to distinguish the pronominal expressions whose antecedents could be denoted by NPs from those that couldn't, a set of labels was developed so that all pronominal expressions in the class of interest could be labeled in some way. This provides material for training/scoring automatic methods of distinguishing anaphors that have nominal antecedents from anaphors of other types and from pronouns that are not anaphoric.

The rest of this chapter is organized as follows. First, an overview of the various choices that must be made will be given. That will be followed by what the current guidelines say with respect to those choices. Finally, a comparison with other coreference guidelines will be given.

3.2 Questions for Coreference Annotation

First, the kind of anaphors and the kinds of antecedent expressions must be established. This choice is inseparable from the choice of the type of anaphoric relation that is being tagged.

For this work we will be considering those pronominal anaphors that depend on entities represented by noun phrases in the text, and the relation is one of identity,

namely that the two expressions, antecedent and anaphor, refer to the same entity or to the same meaning. Discourse, temporal, or event anaphora are not annotated.

Given the choice of type of anaphor and relation, other policy choices follow, in this case with respect to nominal antecedents. A decision must be made whether to annotate just the head of the antecedent expression or to annotate the maximal NP. Related to this is the question of how to annotate NP apposition in expressions like "Bono, the rock star and social activist."

A policy choice must be made on whether or not to tag predicate nominals. Metonymy is a particularly pesky problem to deal with when describing how to recognize when two expressions refer to the "same entity."

3.3 Which Pronominal Anaphors Are Being Annotated

This work only addresses third-person pronouns, demonstratives, and the zero-pronoun ***pro***[1]. The annotation does not include first- or second-person pronouns, quantified expressions such as 一些/"a few", locations such as 这里/"here", or question words such as 谁/"who".

The pronouns that are annotated include: 他/"he,his", 她/"she,her", 它/"it", 其, 之/"its", 他们/"they (masc.)", 她们/"they (fem.)", 它们/"they (neut.)", 此/"this", 这/"this", 这些/"these", 那/"that", 那些/"those", 自己/"oneself", and 彼此/"one another, each other".

When the expressions 这/"this" and 那/"that" (or their plural forms) are used as determiners rather than as demonstrative pronouns they are not tagged. In the CTB parses this is obvious from the part-of-speech (POS) tag (terminal label PN for the pronouns, terminal label DT for the determiners).

[1]Note that the word "pronouns" in this document may occasionally be used broadly to mean not only pronouns like 她/"she" or 他们/"they," but also demonstratives like 此/"this" and 那/"that," as well as the zero pronoun ***pro***.

The Appendix to this chapter (Section 3.10) contains tables listing which words tagged as pronouns by the CTB (as defined by the POS tag "PN") are annotated for this task and which are not annotated.

Possessive pronouns are also annotated, although, due to the bracketing conventions of the CTB, together with the annotation conventions that are used here (described in Section 3.5.2 below), it turns out that the annotation of possessives requires no special treatment.

These examples illustrate how possessive pronouns appear in the CTB.

From **chtb_207** "her emerald necklace":

(NP-OBJ(DNP(NP#1 (PN 她))	she/her
(DEG 的))	DE (poss)
(NP (NN 翡翠)(NN 项链))))))	emerald necklace

From **chtb_249** "its own road":

(NP-OBJ(DNP(NP#2 (PN 自己))	itself
(DEG 的))	DE (poss)
(NP (NN 路))))))	road

As mentioned at the beginning of the chapter, an important reason for using the Chinese Treebank was the realization of zero pronouns by the string ***pro***. Thus all ***pro*** are annotated with an index or a category. Note that "big PRO", denoted by ***PRO*** in the CTB, is *not* annotated[2].

[2]According to the stated use of the string "*pro*" as set forth in the *Bracketing Guidelines* for the CTB [52], ***pro*** is used to represent argument drop, and therefore should be able to be replaced with an overt string. As it is actually used in the CTB, however, ***pro*** does not always appear in a position that permits a lexicalized expression. Since there is no single diagnostic that distinguishes these ***pro*** from the ***pro*** that can be lexicalized, the coreference annotation guidelines instruct that all ***pro*** be coindexed or classified.

3.4 The Types of Anaphor-Antecedent Relations That Are Marked

The term "anaphor" as it is used in anaphora resolution refers broadly to many kinds of expressions that depend for their interpretation on other referring expressions. The annotation here only marks, by coindexing, anaphor-antecedent identity relations when the antecedents are nominal expressions that denote entities in the world. Thus there is a restriction that the syntactic form of the antecedent expression is an NP, or in some cases a QP (when the head noun being quantified is understood)[3].

For example, we are not assigning index numbers to pronouns that refer to events or propositions. Instead there is a special tag (**#DD**) that is used to label such anaphors (to be discussed in Section 3.7 below along with the labels for other categories of pronouns that do not have nominal antecedents).

An example from **chtb_207** follows, illustrating both a coindexable and a non-coindexable pronoun (the parse that follows is taken from the CTB). The demonstrative pronoun 这/zhe4/*this* that is labeled #DD does not refer to an entity denoted by a nominal expression, but to the proposition that Mount Fuji and Tokyo are separated from one another by only 80 km, the first clause (an IP in the parse). It therefore is not assigned an index, but is marked as discourse deictic (**#DD**). In contrast, the pronoun 她/ta1/*she* on the last line refers to the NP 这个岛国/zhe4gedao3guo2/*this island nation*, and is coindexed with that NP.

The attentive reader will ask why the NPs for Mount Fuji and for "this island nation's highest mountain" were not coindexed in the example. The reason is that the text did not contain an anaphor that referred to Mount Fuji, and doing general NP coreference was beyond the scope of this work.

There are a number of questions that must be answered to more fully define the kinds of NP antecedent expressions that will be coindexed with pronominal anaphors.

[3]In the CTB, QPs, quantifier phrases, are expressions of quantity that are equivalent to NPs with respect to their syntactic role.

富士山 离 东京 只 有 80 公里 ———
Mt. Fuji separated from Tokyo only has 80 km –
Mount Fuji is only 80 kilometers from Tokyo –
这#DD 显然 又 是 造化 的 特意 安排 :
this obviously also be Mother Nature DE (poss) on purpose arrangement :
this obviously was also a deliberate arrangement of Mother Nature:
[这 个 岛国]#1 最 高 的 山 与
this M island nation most high DE (mod) mountain and
this island nation's highest mountain and
她#1 的 最 大 的 城市 相依为伴 , 互 为 辉映 。
she DE (poss) most large DE city depend on one another as partners , mutually be reflect .
her largest city depend on one another as partners and reflect one another.

Example from **chtb_207**

These will be discussed after a brief aside to discuss the mechanics of the annotation.

For example, in addition to anaphor-antecedent relations in which the antecedent is referential, as in **#1** above, there are also coindexed anaphor-antecedent relations in which the antecedent is not strictly referential, in a linguistic sense. These will be discussed in Section 3.6.3

Caveat: Even though it is not always strictly linguistically accurate, from now on the terms "corefer", "coreference", and "coreferential" will be used for those anaphor-antecedent relations that have antecedents denoting nominal entities, even when the nominal and pronominal expressions themselves are not strictly referential.

```
 1((IP (IP (IP (NP-PN-SBJ (NR 富士山))                                  Mount Fuji
2               (VP (PP-LOC (P 离)                                  separated from
3                           (NP-PN (NR 东京)))                          Tokyo
4                   (ADVP (AD 只))                                      only
5                   (VP (VE 有)                                         has
6                       (QP-OBJ (CD 8 0)                                80
7                               (CLP (M 公里))))))                      km
8           (PRN (PU ---)
9                 (IP (NP#DD-SBJ (PN 这))                          deictic this
10                    (VP (ADVP (AD 显然))                              clearly
11                        (ADVP (AD 又))                                also
12                        (VP (VC 是)                                   was
13                            (NP-PRD (DNP (NP (NN 造化))           good fortune
14                                         (DEG 的))                DE (poss)
15                            (ADJP (JJ 特意))                        special
16                            (NP (NN 安排))))))))                arrangement
17      (PU : )
18      (IP (NP-SBJ (NP (NP#1 (DP (DT 这)                      this (determiner)
19                                (CLP (M 个)))                         MW
20                            (NP (NN 岛国)))                   island nation
21                        (CP (WHNP-1 (-NONE- *OP*))
22                            (CP (IP (NP-SBJ (-NONE- *T*-1))
23                                    (VP (ADVP (AD 最))              most
24                                        (VP (VA 高))))              high
25                                (DEC 的)))                      DE (mod)
26                        (NP (NN 山)))                            mountain
27                    (CC 与)                                          and
28                    (NP (DNP (NP#1 (PN 她))                          her
29                                   (DEG 的))                    DE (poss)
30                        (CP (WHNP-2 (-NONE- *OP*))
31                            (CP (IP (NP-SBJ (-NONE- *T*-2))
32                                    (VP (ADVP (AD 最))              most
33                                        (VP (VA 大))))              big
34                                (DEC 的)))                      DE (mod)
35                        (NP (NN 城市))))                         city
36                    ...
```

(Line numbers are added for discussion purposes, but do not correspond to the line numbers in the actual CTB parse).

3.5 The Mechanics of the Annotation

3.5.1 Annotation framework/environment

There are two basic approaches to annotation: annotating source files directly or stand-off annotation that does not alter the source data, but instead uses an annotation tool such as WordFreak[4] to keep track of annotations using pointers into the source files.

The potential downsides to standoff annotation systems are flexibility, portability, and maintenance.

Annotation directly on source files (as is the case for the different parts of the CTB) has the virtue of simplicity, and makes the annotations easily available for general use independent of the need for a particular interface (API) to interpret the annotation files. One downside is that the "users" of the data must devise their own interfaces.

The annotation here was done directly on copies of the CTB parsed files using emacs. Annotation was done by native Mandarin speakers. Adjudication was done by committee.

3.5.2 Annotation conventions

Anaphors were tagged with indices when they were coreferential with nominal entities, and tagged with category labels when they were not.

Annotators were asked to annotate every pronominal expression from the list discussed above in Section 3.3 (a complete list is given in a table in 3.10) that was found in the file. If the pronoun was coreferential, the relation was indicated by putting the same index number on the anaphoric NP and on the antecedent NP. When the pronouns were not coindexable the category label was put on the

[4] **http://sourceforge.net/projects/wordfreak/**

pronominal NP only[5].

The index for an entity is annotated using #<**index-number**>. A particular <**index-number**> is appended to the NP label immediately projected by the anaphoric expression, i.e., the immediate parent NP of the (**PN pronoun**) or the (**-NONE- *pro***). Then the same index is appended to the NP label of the antecedent's expression. Indices were *not* placed on terminal words. For example:

(NP#1-SBJ (PN 他))	NOT: (NP-SBJ#1 (PN 他))
(NP#1-SBJ (PN 他))	NOT: (NP-SBJ (PN#1 他))
(NP#1-SBJ (-NONE- *pro*))	NOT: (NP-SBJ (-NONE- *pro*#1))

Each unique entity referred to in a file received a single, file-unique index.

Annotators were asked to do a bit more work than to just coindex the immediate antecedent of each anaphor. Once an entity had been "selected" by a pronominal anaphor, annotators were also asked to put that entity's index on every referential expression in the file that referred to that unique semantic entity.

From a strictly linguistic point of view, only the most recent meaning-bearing antecedent expression should be tagged for one particular anaphoric relation. Annotating additional referential NPs, especially *after* the last anaphor, does not make sense.

The choice to put a referent's index on all coreferential expressions (in the strict sense of coreferential) in the same text was made in order to extend the usefulness of this annotated corpus. Because the primary goal of annotation was pronominal anaphora resolution and not general noun phrase coreference, however, the annotators were only asked to find the complete chains for those entities that were antecedents of at least one pronominal anaphor.

[5]even in the case of anaphors tagged with **#DD**

3.6 Policy Choices

3.6.1 Nesting level and appositives

When annotating nominal expressions, there is always a choice between annotating the entire extent of a noun phrase *vs.* just annotating the head of the noun phrase. In this annotation scheme, annotators were asked to annotate the maximal NP, that is to annotate the top-most NP parent of the head NP.

Part of the motivation for this decision was that it is straightforward to determine the head of a maximal NP, but the reverse is not necessarily true. In addition, in this domain, different entities frequently have identical heads, and it was found to be clearer to annotate the full NP that gave the descriptive material that distinguished expressions with identical heads.

An unanticipated side-effect of this early decision was that it meant the implementation of the Hobbs algorithm did not require any adjustment due to the difference in the locations of nominal heads between English (on the left) and Chinese (on the right). This will be discussed in Chapter 4.

The following example from **chtb_002** illustrates one reason for annotating the parent-level NP rather than the lowest-level NP.

The first ***pro***, which is marked **NP#2**, refers to an entity that appeared earlier in the text, namely 外商投资企业/wai4shang1 tou2zi1 qi3ye4/*foreign-owned enterprises*.

The head noun phrase at the lower level, 进出口总值/jin4chu1kou3 zong3zhi2/*the total value of imports and exports*, is the same in both **NP#3** and **NP#4**, but semantically **NP#3** is a subset or part of **NP#4**.

[*pro*#2	去年	实现	进出口	总值] #3	达	一千零九十八点二亿	美元	，
ϕ #2	last year	realize	imports	total value	reach up to	1098.2 x 10^8	US$	,

The total value of imports and exports that they realized last year reached 109.82 billion US$,

pro #3	占	[全	国	进出口	总值] #4	的	比重
ϕ #3	stand	entire	country	imports	total value	DE (mod)	ratio

the proportion that this represented in the value of imports and exports of the entire country

提高	到	百分之三十九	。
rise up	reach	39%	.

rose up to 39 percent.

Example from **chtb_002**

3.6.2 Appositives

The rule of annotating high means that appositive constructions do not have indices on some referring expressions that could independently, uniquely identify their referents in the context of the given file. For example, in **chtb_249**, an article from August, 1994, in the noun phrase "the Israeli Foreign Minister, Simon Peres", either expression denotes the same individual in the world, but only the entire phrase would receive an index[6].

3.6.3 Bound anaphora

As mentioned in Section 3.4, there are pronominal anaphors that have non-referring nominal antecedents. For example, in

"every man who knows his own mind"

"every man" does not refer to an individual entity in the world, but denotes a class of "men".

[6]This was changed from the first version of the guidelines ([7]) in which *only* the two lower NPs were labeled, and the parent NP was not, but only in cases of apposition in which a name was paired with a definite description or title. Annotating appositives at the parent level instead was easier for the annotators as well as being more consistent.

chtb_002

(IP (**NP#3**-SBJ	entity **#3**
(CP (WHNP-1 (-NONE- *OP*))	
(IP (**NP#2**-SBJ (-NONE- ***pro***))	entity **#2**
(VP (NP-TMP (NT 去年))	last year
(VP (VV 实现)	realize
(NP-OBJ (-NONE- *T*-1))))))	
(NP (NN 进出口)	imports+exports
(NN 总值)))	total value
(VP (VV 达)	reach
(QP-OBJ (CD 一千零九十八点二亿)	1098.2 x 10^8
(CLP (M 美元)))))	US$
(PU ，)	
(IP (NP-SBJ (CP (WHNP-2 (-NONE- *OP*))	
(CP (IP (**NP#3**-SBJ (-NONE- ***pro***))	entity **#3**
(VP (VV 占)	stand
(**NP#4**-OBJ (NP (DP (DT 全))	entire
(NP (NN 国)))	country
(NP (NN 进出口)	imports+exports
(NN 总值)))	total value
(NP-EXT (-NONE- *T*-2))))	
(DEC 的)))	DE (mod)
(NP (NN 比重)))	proportion
(VP (VRD (VV 提高)(VV 到))	rise up to
(QP-OBJ (CD 百分之三十九)))))	39 percent
(PU 。)))	

In these cases of bound anaphora the pronoun "his" would be coindexed with the antecedent expression "every man", but if "every man" appeared again in the text it would not receive an index. If the second "every man" had its *own* bound anaphor, then that anaphor-antecedent pair would receive a new index.

3.6.4 NP predicates

When an entity that has been selected by an anaphoric relation with a pronoun appears as the subject of a copula followed by an NP predicate, the NP predicate is also tagged with the entity's index *if and only if* the NP-PRD expression refers *uniquely* to the same entity.

This passage from **chtb_007** illustrates a coindexable NP-PRD and an NP-PRD that should not be coindexed because it does not uniquely identify a single entity.

崇明 **#2** 是 [中国 第三 大 岛] **#2** ，
Chongming is China third large island ,
Chongming *is China's third largest island and*
具有 优越 的 地理 条件 和 悠久 的 历史 ，
possess excellent DE geography qualification and long DE history
possesses excellent geographical qualifications and a long history.
改革 开放 以来 ， 崇明县 **#2** 的 经济 建设
reform opening since , Chongming County DE economy construction
Since reform and opening up, **Chongming County***'s development in economic construction*
和 对 外 开放 发展 迅猛 ， ...
and toward outside opening development swift , ...
up vigorous , ...
and opening up to the outside has been swift and vigorous, and ...
pro 是 中国 综合 实力 百 强 县 之一 。
ϕ **#2** is China constitute strength 100 strong county oneof .
it *is one of the hundred strong counties that constitute China's actual strength.*

Example from **chtb_007**

The NP-PRD at lines 3-5 in the parse on page 41,

```
 1((IP(IP (NP#2-PN-SBJ (NR 崇明))                                  Chongming#2
2         (VP(VP (VC 是)                                                 is
3                 (NP#2-PRD (NP-PN (NR 中国))                          China
4                           (QP (OD 第三))(ADJP (JJ 大))          third large
5                           (NP (NN 岛))))                             island
6            (PU ，)
7            (VP (VV 具有)                                           possesses
8                (NP-OBJ(NP(CP(WHNP-1 (-NONE- *OP*))
9                              (CP(IP(NP-SBJ (-NONE- *T*-1))
10                                 (VP (VA 优越)))                   superior
11                             (DEC 的)))                          DE (mod)
12                          (NP(NN 地理)(NN 条件)))  geographic qualifications
13                     (CC 和)                                          and
14                     (NP(CP(WHNP-2 (-NONE- *OP*))
15                            (CP(IP(NP-SBJ (-NONE- *T*-2))
16                               (VP (VA 悠久)))             long-standing
17                            (DEC 的)))                          DE (mod)
18                          (NP (NN 历史)))))))                      history
19   (PU ，)
20   (IP(LCP-TMP(NP(NN 改革)(NN 开放))                  reform and opening up
21              (LC 以来))                                            since
22      (PU ，)
23      (IP(IP(NP-TPC(DNP(NP#2-PN (NR 崇明县))    Chongming County #2
24                        (DEG 的))                                 DE (poss)
25                   (NP(NP(NN 经济)(NN 建设))       economic construction
26                      (CC 和)                                        and
27                      (NP(PP(P 对)(NP(NN 外)))          toward outside
28                         (NN 开放))))                          opening up
29            (NP-SBJ (NN 发展))                                  development
30            (VP (VA 迅猛)))                                swift and vigorous
31         (PU ，)
32         ...
33         (IP(NP#2-SBJ (-NONE- **pro**))                              #2
34            (VP(VC 是)                                               is
35               (NP-PRD(NP (NR 中国))                              China
36                      (NP(ADJP (JJ 综合))                      make up
37                         (NP (NN 实力)))                       strength
38                      (NP (QP (CD 百))                            100
39                         (ADJP (JJ 强))(NP (NN 县)))       strong county
40                      (NP (NN 之一)))))))                      one of
```

中国 第三 大 岛/zhong1guo2 di4san1 da4 dao3/*China's third largest island*, is Chongming and no other.

The NP-PRD on lines 35-40, 中国综合实力百强县之一/*one of the hundred strong counties that constitute China's actual strength*, however, is not a definite reference and could refer to any one of many other counties (or potentially 99 other counties if one takes the words literally) as well as to Chongming.

A question test was used to distinguish when to annotate an NP-PRD. If the question formed by rewording the NP-PRD had a unique answer, then the expression could be coindexed.

For example, for the NP-PRD on lines (3-6) in **chtb_007** one could ask: 中国第三大岛是什么地方?/*What place is China's third largest island?* and the answer would uniquely be "崇明/Chongming".

This diagnostic also applies to a small number of NP-OBJ expressions.

There are certain verbs other than the copular verbs 是/shi or 为/wei for which the NP object is the same entity as the subject entity, following a change of state. These verbs are words like 成为/cheng2wei2/*become* or 建成/jian4cheng2/*establish*. In these cases the NP-OBJ should be indexed as well as the subject NP, since the verb has established that the second description picks out the same identity.

This example is from **chtb_014**, in which an anaphor elsewhere in the file has China as its antecedent:

中国	#	已	[成为	韩国	最	大	的	投资	对象国]	#2	。
China		already	become	Korea	most	big	DE	invest	target country		.

China has already become Korea's largest target country for investment.

In other words, the answer to the question 韩国的最大的投资对象国是谁?/*Who is Korea's largest target country for investment?* is uniquely 中国/*China*.

As far as the text in **chtb_014** is concerned, both expressions denote the same entity in the world.

chtb_014

(IP (NP#**2**-PN-SBJ (NR 中国))	China
(VP (ADVP (AD 已))	already
(VP (VV 成为)	become
(**NP#2**-OBJ (NP-PN (NR 韩国))	Korea
(CP (WHNP-2 (-NONE- *OP*))	
(CP (IP (NP-SBJ (-NONE- *T*-2))	
(VP (ADVP (AD 最))	most
(VP (VA 大))))	large
(DEC 的)))	DE (mod)
(NP (NN 投资)	investment
(NN 对象国))))))	target country

3.6.5 Split antecedents

The battle-scarred reader will have noticed that among the list of pronominal expressions to annotate there are plural and "mutual" expressions.

Sometimes life is easy and a plural pronoun refers to a simple plural expression:

```
[The girls]#1  went to space camp last summer.
[Their]#1  moms were all envious.
```

In such cases, both plural expressions can be tagged with a single index.

The antecedents for a plural expression can also appear as references to two or more individuals in the preceding context, as in:

```
[Sally and Jill]#2  wanted to go to space camp,
 but [they]#2  applied too late to get in.
```

In this example there is still a single (albeit conjoined) constituent NP, "Sally and Jill", that may be coindexed with "they".

But there are situations in which a single index will not work.

```
[Sally]#1 , wanted to go to space camp.
```

```
When [her]#1  younger sister [Jill]#2  found out,
   [she]#2  wanted to go too.
But [their]#? parents couldn't afford it that summer.
```

To handle these split antecedents, the convention was used to assign each individual an index, and then to put multiple indices on the plural anaphoric expression.

```
But [their]#1#2 parents couldn't afford it that summer.
```

Split antecedents can be quite complicated. In an article discussing the birth of sextuplets, there were different anaphors whose antecedent expressions corresponded to different subsets of the six individual infants (starting with the two that died and the four survivors). Using six indices for the individual infants and multiple indices on the various plural expressions made the tagging straightforward, if tedious.

3.6.6 Constituency problems

Once an entity had been picked out as the antecedent for a pronominal anaphor, the annotators were asked to coindex every NP expression in the file that referred to the same, unique entity. Sometimes it was not possible to annotate an expression denoting the entity due to the constituent bracketing in the CTB parse. Typically these cases were found in files that were discussing bilateral relations between two countries or organizations, and the entity being tagged was one of the parties.

For example a file might have an anaphor referring to 中国/China, which had the index **#2**, and there was another anaphor referring to 美国/America and it had been assigned the index **#3**. It would not be uncommon for there to be the expression (**NP** (**NR** 中/China)(**NR** 美/America)) somewhere in that file. Ideally, 中/China should be tagged with **#2** and 美/America with **#3**, but neither projects its own NP, and the annotating conventions do not allow putting indices or labels on parts of speech.

Not being able to label China and America is a limitation of the direct annotation methodology and the convention that only phrase-level expressions are tagged (Section 3.5.1).

These cases were called "constituency problems," and annotators were asked to make a note in their comments files to document them, but no further action was taken.

There was no case in which the immediate antecedent of an anaphor had such a "constituency problem."

3.7 The Categories for Those Anaphors That Cannot Be Coindexed

As mentioned above, anaphors that have antecedent expressions that denote nominal entities or types ("coindexable" pronouns) make up just one category of pronominal usage. The remaining pronouns were assigned tags for the following categories:

1. **discourse deictic** (**#DD**)

2. **existential** (**#EXT**)

3. **inferrable** (**#INFR**)

4. **ambiguous** between two (or more) possible antecedents, according to different possible interpretations of the text (**#AMB**)

5. **arbitrary reference** (**#ARB**)

6. **unknown** (**#UNK**)

3.7.1 Discourse Deictic (#DD)

Pronouns that refer to propositions, situations, or events in the preceding text are annotated with the label **#DD**. Syntactically, the antecedent for a discourse deictic pronoun is typically an IP. An example of this from **chtb_207** was given in Section 3.4 above (page 34), and the relevant portion of the parse is repeated here:

```
chtb_207 1((IP (IP (IP (NP-PN-SBJ (NR 富士山))                          Mount Fuji
2                      (VP (PP-LOC (P 离)                          separated from
3                                  (NP-PN (NR 东京)))                  Tokyo
4                          (ADVP (AD 只))                              only
5                          (VP (VE 有)                                 has
6                              (QP-OBJ (CD 8 0)                        80
7                                      (CLP (M 公里))))))              km
8                  (PRN (PU ---)
9                       (IP (NP#DD-SBJ (PN 这))                    deictic this
10                         (VP (ADVP (AD 显然))                        clearly
11                             (ADVP (AD 又))                          also
12                             (VP (VC 是)                             was
13                                 (NP-PRD (DNP (NP (NN 造化))) good fortune
14                                                 (DEG 的))     DE (poss)
15                                 (ADJP (JJ 特意))                    special
16                                 (NP (NN 安排))))))))            arrangement
                   ...
```

Note that nominalizations of events are NPs, and are therefore perfectly reasonable antecedents.

3.7.2 Existential (#EXT)

As a rule, when the zero pronoun ***pro*** appears as the subject of the existential verb (VE) 有/you3/*to have* (or the negative existentials 没有/mei2you3/*not have* or 无/wu2/*without*) it is not referring to any particular entity, but is being used in a way similar to the English existential "there are" or "there is".

When ***pro*** is used in this way, the NP it projects is annotated **#EXT**.

Here is an example from **chtb_001**:

建筑	公司	进区	，					
construction	company	enter region	,					
[When] a construction company enters the region,								
有关	部门	先	送上	这些	法规性	文件	，	
appropriate	bureau	first	deliver	these	regulatory	documents	,	
the appropriate bureau first delivers these regulatory documents,								
pro **#EXT**	然后	有	专门	队伍	进行	监督	检查	。
ϕ	afterwards	have	expert	team	carry out	supervision	inspection	.
and afterwards **there is** *a contingent of experts that carries out a supervisory inspection.*								

Example from **chtb_001**

chtb_001

```
(IP (NP#EXT-SBJ (-NONE- *pro*))                                  #EXT
    (VP (ADVP (AD 然后))                                         afterwards
        (VP (VE 有)                                              to have
            (IP-OBJ (NP-SBJ (ADJP (JJ 专门))                     specialist
                            (NP (NN 队伍)))                      contingent
                    (VP (VV 进行)                                carry out
                        (NP-OBJ (NN 监督)                        supervise
                                (NN 检查)))))))                  inspection
```

Note that this category may appear with the verb negated, as this example from **chtb_249** illustrates:

3.7.3 Inferrable (#INFR)

There are cases in which the entity to which the pronoun refers is not represented directly by a string that appears in the text, but must be inferred from entities that are denoted by some noun phrase, or inferred from world knowledge about the

pro#**EXT**	假如	没有	这些	财富	，			
ϕ	if	not	these	riches	,			
If it were not for these riches,								
人们	将	不	可能	描述	历史	的	进程	。
people	will	not	be able	describe	history	DE	progress	.
	(fut)		to			(poss)		
people would not be able to describe the course of history.								

Example from **chtb_249**

chtb_249

((IP (IP-CND (**NP#EXT**-SBJ (-NONE- *pro*))	**#EXT**
(VP(ADVP (CS 假如))	if
(VP(VE 没有)	not have
(NP-OBJ(DP (DT 这些))	these
(NP (NN 财富))))))	riches
(PU ，)	
(NP-SBJ (NN 人们))	people
(VP (ADVP (AD 将))	will/would
(ADVP (AD 不))	not
(VP (VV 可能)	be able to
(VP (VV 描述)	describe
(NP-OBJ(DNP(NP (NN 历史))	history
(DEG 的))	DE (poss)
(NP (NN 进程))))))	course
(PU 。)	

context. In these situations, the annotators are asked to mark the anaphor with **#INFR**, and to write in the comments file what specific entity the anaphor refers to.

In this example from **chtb_072**, the ***pro*** refers to the amount of exports of the state-owned enterprises in Xiamen, which is inferred from the previous IP:

今年	前	十	个	月	厦门市	国有	企业
this year	previous	10	M	month	Xiamen municipality	state-owned	enterprises

In the previous ten months of this year, Xiamen's state-owned enterprises

出口	十六点三七亿	美元	，
export	16.37 x 10^8	US$	.

exported 1.637 billion US$,

pro	**#INFR**	增长	百分之二十三点六	，
ϕ		increase	23.6%	

increasing by 23.6 percent,

并	呈现	出	较	强	的	增长	后劲	。
moreover	appear	out	comparatively	strong	DE	growth	stamina	.

and showing relatively strong growth stamina.

3.7.4 Ambiguity between possible referents in the text (#AMB)

There are places in texts in any language in which there can be two (or sometimes more than two) possible interpretations of a sentence containing a pronoun. The category **#AMB** was used when the antecedent of the pronominal anaphor had more than one possibility.

For example, in the first sentence in the file **chtb_100**, it is not clear what the subject of the verb 发现/fa1xian4/*discover* is. It could be (中国)陆上石油工业/(zhong1guo2)lu4shang4 shi2you2 gong1ye4/*(China's) land-based petroleum industry*, some responsible people who work in the industry, or perhaps even just 中国/*China*. Given the headline:

The first sentence of the file is:

chtb_072

((IP(IP(NP-TMP (NP (NT 今年))	this year
(DP (DT 前)	previous
(QP (CD 十)(CLP (M 个))))	10 MW
(NP (NN 月)))	months
(NP-SBJ (NP-PN (NR 厦门市))	Xiamen Municipality
(ADJP (JJ 国有))	state-owned
(NP (NP (NN 企业))))	enterprises
(VP (VV 出口)	export
(QP-OBJ (CD 十六点三七亿)	16.37 x 10^8
(CLP (M 美元)))))	US dollars
(PU ，)	
(IP (NP#**INFR**-SBJ (-NONE- ***pro***))	**#INFR**
(VP (VP (VV 增长)	increase
(QP-EXT (CD 百分之二十三点六)))	23.6 percent

...

中国	去年	发现	十	个	亿	吨	级
China	last year	discover	ten	M	10^8	ton	level
储量	规模	的	油气区				
reserve	scale	DE	oil-gas field				

HL:*Last Year China Discovered 10 Oil Fields With 100 Million Ton Reserves*

中国	陆上	石油	工业	在	过去	一	年	中
China	land-based	oil	industry	in	past	one	year	middle
取得	重大	成绩	：					
achieve	great	success	:					

China's land-based petroleum industry achieved great success in the past year:

pro **#AMB**	全	年	发现	十	个	亿	吨	级
ϕ **#AMB**	entire	year	discover	ten	M	10^8	ton	level
储量	规模	的	油气区	。				
reserve	scale	DE	oil-gas	.				
		(mod)	field					

ten oil-gas fields with 100 million ton reserves were discovered in the whole year.

chtb_100

((IP (IP (NP-SBJ (NP-PN (NR 中国))	China
(NP (NN 陆上)	land-based
(NN 石油)(NN 工业)))	petroleum industry
(VP (PP-LOC (P 在)	in
(LCP (QP (NP (NT 过去))	last year
(QP (CD 一)(CLP (M 年))))	one year
(LC 中)))	middle
(VP (VV 取得)	obtain
(NP-OBJ (ADJP (JJ 重大))	great
(NP (NN 成绩))))))	success
(PU ：)	
(IP (**NP#AMB**-SBJ (-NONE- *pro*))	**#AMB**
(VP (DP-ADV (DT 全)(CLP (M 年)))	entire year
(VP (VV 发现)	discover
(NP-OBJ (QP (CD 十)(CLP (M 个)))	ten MW
(DNP (NP (NP (QP(CD 亿)	10^8
(CLP (M 吨)))	ton
(NP (NN 级)))	level
(NP (NN 储量)	reserves
(NN 规模)))	scale
(DEG 的))	DE (mod)
(NP (NN 油气区))))))	oil gas field
(PU 。)))	

The category **#AMB** is different from **#INFR** in that for **#INFR** it should be clear to the reader what the single referent is, but there is no expression in the text that directly denotes the desired antecedent.

3.7.5 Arbitrary Reference (#ARB)

There are cases in which ***pro*** does not refer to a single, specific entity that is named in the text, nor does it refer to any specific entity that can be inferred from the text. In these non-specific instances, sometimes the ***pro*** stands for anyone, or people in general, and it would be possible for the ***pro*** to be replaced by an overt, generic expression, such as 某人/mou3ren2/*someone* or 我们/wo3men//*we*, in the general sense[7].

The **#ARB** in the following example from **chtb_027** illustrates one such situation.

据	*pro* **#ARB**	认为	,	此	次	访问	的	目的
according to	ϕ	to think that	,	this	time	visit	DE	goal

It is thought that the purpose of this visit

是	为了	改善	和	发展	两国	关系	,
be	for	improve	and	develop	two countries	relations	,

is to improve and develop the relationship between the two countries,

That is, the ***pro*** does not refer to any specific person but to people in general. The most natural English translation would be "it is understood". There is some general, agentive entity that is the subject of the verb.

At other times, the ***pro*** cannot be interpreted to refer to even a general, anonymous agent. Instead, the attention of the phrase is on the action denoted by the verb itself, rather than on any subject of the verb. No overt subject could replace the ***pro*** in these situations.

[7] "We" in Chinese plays the role that the catch-all "you" does in English.

chtb_027

((IP (PP(P 据)	according to
(IP(**NP#ARB**-SBJ (-NONE- *pro*))	**#ARB**
(VP (VV 认为))))	think/consider/hold
(PU ，)	
(NP-SBJ (DNP(NP(DP(DT 此)	this
(CLP (M 次)))	M-time
(NP (NN 访问)))	visit
(DEG 的))	DE (pos)
(NP (NN 目的)))	goal
(VP (VC 是)	is
(PP-PRD(P 为了)	in order to
(IP(NP-SBJ (-NONE- *PRO*))	*PRO*
(VP(VP(VP(VV 改善)	improve
(NP-OBJ (-NONE- *RNR*-1)))	
(CC 和)	and
(VP(VV 发展)	develop
(NP-OBJ-1(NP(QP (CD 两))	two
(NP (NN 国)))	countries
(NP (NN 关系)))))	relations
(PU ，)	
(VP (... yadda yadda yadda ....)	

3.7.6 Unknown (#UNK)

The label **#UNK** was used only as a trap for debugging the guidelines, and does not appear in the final annotated files.

3.8 A Comparison of This Annotation Scheme With Other Annotation Efforts

As stated above, the primary focus of the annotation effort described in this document is to provide "gold-standard" data for training and evaluating methods for resolving pronominal anaphors that have nominal antecedents.

In addition to the work necessary to provide data for this narrow task, the annotation effort was extended to categorize those pronominal expressions that were *not* nominal-dependent anaphors. A further extension to the annotation was to mark all coreference relations for those specific NP entities that served as antecedents to pronominal anaphors.

While the bias of the annotation is on the computational side of computational linguistics, the intent was to try to make as linguistically motivated choices as possible.

This annotation effort differs in scope and priorities from some other coreference annotation work. Two other annotation schemes will be discussed here, those from the Message Understanding Conference (MUC) Coreference Task [18] and those for the 2004 Chinese Automatic Content Extraction (ACE) bakeoff[8].

The next sections will discuss some key differences between the guidelines written for those two tasks and our guidelines for the CTB.

[8]Under the heading *ACE3* under **http://projects.ldc.upenn.edu/ace/annotation/previous/**.

3.8.1 Goals

The MUC-7 Coreference Task Definition's priorities are clearly stated at the beginning of the guidelines:

1. Support for the MUC information extraction task
2. Ability to achieve good (ca. 95%) interannotator agreement
3. Ability to mark text up quickly
4. Desire to create a corpus for research on coreference and discourse phenomena, independent of the MUC extraction task

In contrast to having a specific coreference task, the ACE effort is focused on entity detection and tracking. The "coreference" task within ACE is thus to detect expressions, called "mentions", that either identify or that describe the same entity and to give those mentions the same entity ID number. The expressions are not necessarily strictly referential.

There is no single ACE document dedicated to the problem of coreference *per se*. Instead the *Entity Link Tracking* and *Entity Detection and Tracking* guidelines may be taken together to provide the information for the following comparisons.

3.8.2 IDENT relation

MUC

The priorities of MUC led to a decision to only annotate identity (IDENT) coreference relations. That is, expressions A and B are only coreferent if A and B both denote the same referent. Because this IDENT relation is symmetric and transitive, all expressions that refer to a single entity form an equivalence class defined by the IDENT relation.

The ACE relation being annotated was effectively also the IDENT relation.

The CTB annotation also used the IDENT relation, but with two implementation differences from MUC and ACE. The first difference was in which entities would be selected for annotation. The second restriction was in the interpretation of what type of NP would be considered "identical". These will be discussed below.

3.8.3 Pronouns and antecedents: "markables"

MUC calls the kinds of expressions that can be in the IDENT relation "markables." The set of markables in a MUC text is far more inclusive than in our annotation of the CTB. All noun phrases, nouns, and pronouns were potential "markables" in MUC. Noun phrases included dates, currency expressions, and percentages.

In our annotation only those noun phrases that served as antecedents of pronominal anaphors were tagged. When those antecedent NPs happened to be referential, then all expressions in the same text that referred, unambiguously and uniquely, to the same entity would be coindexed with the original anaphor-antecedent pair. Dates, currency expressions, and percentages were *potential* antecedents for the CTB annotation, although they might have a QP rather than an NP phrase label. Thus the difference between MUC and our annotations in this respect was not so much the type of entity but in how the "markables" were selected.

The more important restriction that was made in the CTB annotations was that *indefinite* NPs were *not* tagged.

In ACE, the set of nominal and "pronominal" expressions that are assigned entity identification numbers (IDs) is restricted to those expressions that describe just seven types of entities: people, organizations, geo-political entities (GPEs), locations, facilities, weapons, and vehicles[9]. As was the case for MUC, indefinite expressions could be tagged with an entity ID.

In MUC, "pronouns" included all cases of personal pronouns as well as demonstratives. In the case of ellipsis "the empty string is not markable". In CTB we did

[9]This is the 2004 list

not tag first and second person pronouns, but did tag the ***pro*** "empty string".

The ACE definition of pronominal is far more inclusive. In addition to the standard set of overt pronouns (first, second, and third person, reflexives, locatives, and demonstratives), ACE labels a number of different types of nominal expressions as "pronominal". Several are in fact headless nominals. They include bare quantifiers (很多/"very many"), one and count anaphora, percentages, headless nominals that end in 的/DE, and expressions of the form XX之一/"one of XX". Zero pronouns are not considered.

Because only expressions that denoted or described attributes of one of the seven entity types of interest were given entity IDs in ACE, and because the primary goal of ACE was entity detection, it could happen that there was only one "mention" of an entity in a document, and there was no coreference at all. This could be true of "pronominal" expressions as well as names and nominals.

It also could be the case in an ACE text that an "ordinary" overt pronoun was *not* tagged because its antecedent was not on the list of entities of interest.

3.8.4 Apposition

Both ACE and MUC include appositional phrases in the extent of the maximal noun phrase that is marked, but also annotate a coreference relation between the descriptive appositional phrase alone and the entire noun phrase, as shown in this example from the MUC guidelines:

```
[Julius Caesar, [a well-known emperor]_1 ]_1
```

ACE calls the descriptive phrases "attributive mentions" and for both ACE and MUC these markable phrases may be indefinite NPs. Appositives are not marked in MUC, however, when they are negative or when they represent just a partial match with the head nominal.

As explained in Section 3.6.2 above, because CTB antecedent expressions are annotated at the top-most NP level, appositives are not marked separately from the

head NPs they modify. Only the maximal extent is tagged.

3.8.5 Bound anaphors

Both MUC and ACE tagged coreference relations between quantified noun phrases and pronouns that depended on them, even though they did not meet the strict linguistic definition of "referential", as this example from MUC illustrates:

```
[Most computational linguists]_1  prefer [their]_1  own parsers.
```

The CTB annotation also tagged these "bound anaphors", but there would be no additional coreferring expressions that would be coindexed.

3.8.6 Predicate nominals

As in MUC, our CTB annotation will coindex NP-PRD expressions when they are coreferent with copular subjects that are antecedents of pronominal anaphors (or coreferent with such antecedents). Unlike MUC, which marks as coreferent indefinite NPs that are predicate nominals, our annotation is restricted to definite NPs, as was discussed in Section 3.6.4 above.

3.8.7 Time-dependent identity

In Section 6.4 of the MUC guidelines it states that:

> "Two markables should be recorded as coreferential if the text asserts them to be coreferential at ANY TIME."

This statement is aimed at capturing situations illustrated by sentences like:

"Henry Higgins, who was formerly sales director for Sudsy Soaps, became president of Dreamy Detergents."

MUC would tag that sentence:

```
[Henry Higgins, who was formerly [sales director for Sudsy Soaps]_1 ]_1,
 became [president of Dreamy Detergents]_1.
```

that is, making both titles coreferent with Henry Higgins.

But what if there are two sentences:

"Henry Higgins, who was formerly sales director for Sudsy Soaps, became president of Dreamy Detergents. Sudsy Soaps named Eliza Dolittle as sales director effective last week."

Henry Higgins was marked as coreferential with both "sales director of Sudsy Soaps" and "president of Dreamy Detergents" according to the "ANY TIME" policy. Clearly "Eliza Dolittle" is coreferent with "sales director" of "Sudsy Soaps". Following the literal transitivity of the IDENT relation to make the connections 'Henry Higgins == SS sales director == Eliza Dolittle', which, however, just does not make sense.

The MUC guidelines therefore state that the equivalence should be broken at the sales director equivalence. The two individual "sales director" mentions, with their two coreferent names would be put into different equivalence classes.

MUC uses the semantic terms "extensional descriptor" ("an enumeration of the member(s) of a set by (unique) names"), "intensional description" ("a predicate that is true of an entity or set of entities"), and "grounding instance in a coreference chain" ("the first extensional description in the chain"), to help elucidate the handling of these time-dependent expressions. The intensional descriptions are basically functions that may have different values at different times. The extensional descriptors are the values of those intensional descriptors in the clause's time frame. The intensional expressions might be viewed as functions that take time (and other context such as location) as an argument.

The problem is that the simple coreference annotations used in MUC, ACE, and CTB do not have the power to express these functions.

This problem in CTB happens most often with some numerical intensional expression like a quota or a growth rate paired with extensional expressions in the form of QP-EXT expressions.

In general, the CTB annotations avoided the semantically nonsensical equivalences by not annotating QP-PRD expressions.

Only when a QP expression was the antecedent of a pronominal anaphor was it tagged, and there was no coreference annotation of that antecedent expression. Numbers/quantities of all kinds were considered values that did not uniquely denote entities in the world.

3.8.8 Metonymy

In MUC, coreference was determined with respect to "coerced entities", that is, interpreting the intended meaning of a metonymic mention instead of the literal one, but different facets of geopolitical entities were considered to stand for the same entity, as in this example:

> "The United States is a democracy. The United States has an area of 3.5 million square miles."

in which both "United States" mentions would be coindexed.

In ACE, the problem was split. At the top level entities such as countries, provinces, states, cities, etc. would be labeled as GPEs. But they have multiple facets, and these facets, such as their location, their population, and their governing bodies would be identified in a "role" label (PER, LOC, ORG, or GPE). Mentions with different role labels still shared the same entity ID, however, so in effect, as in MUC, there was coreference between different "roles" of a GPE.

In the CTB annotation, the annotators were instructed to try to distinguish places from governments when the context and the verb made such distinctions clear. Sometimes this was easy, sometimes this was not easy. The MUC/ACE approach would lead to greater inter-annotator agreement.

3.9 Summary

- Goal
  - CTB: Developing and training pronominal coreference resolvers
  - MUC: Information Extraction (IE)
  - ACE: Entity Detection and Tracking (EDT)
- Pronominal Expressions
  - CTB: zeroes, third-person, demonstratives, reflexive, possessives
  - MUC: no zeroes, personal pronouns (all cases and possessive), demonstrative pronouns
  - ACE: no zeroes, one- and count anaphora, headless nominals, bare quantifiers, percentages, <mod>-DE expressions, "one of X"
- Types of possible antecedents
  - CTB: any definite NP, and all NPs that corefer with a coindexed pronoun
  - MUC: nouns and NPs
  - ACE: only expressions naming or describing persons, organizations, locations, geopolitical entities, facilities, weapons, or vehicles
- Relation
  - CTB: IDENT strict
  - MUC: IDENT with indefinites
  - ACE: IDENT with indefinites and generics
- Bound anaphors
  - CTB: yes

  - MUC: yes
  - ACE: yes

- Predicate nominals
  - CTB: only if unique
  - MUC: yes including indefinites
  - ACE: yes including indefinites

- Apposition
  - CTB: annotate high, appositive phrase not tagged
  - MUC: both maximal extent and appositive tagged same index
  - ACE: both maximal extent and appositive tagged same index

- Metonymy
  - CTB: attempt to distinguish roles of GPEs
  - MUC: all facets of GPE coindexed
  - ACE: all facets of GPE coindexed

- Time-dependent identity
  - CTB: handled on a case-by-case basis
  - MUC: impossible chains broken at common title
  - ACE: doesn't say

3.10 Appendix:

Table 3.1: **A. List of words that have the tag PN in CTB that should be annotated:**

本人	oneself ; in person ; I
本身	itself
彼此	each other ; one another
此	this
那	that
其	he/she/it/they ; his/hers/its/theirs
他	he
他们	they (he)
它	it
它们	they (it)
她	she
她们	they (she)
这	this
这些	these
之	it ; its
自己	oneself
自家	oneself
自身	self ; oneself

Table 3.2: **B. List of words that have the tag PN in the CTB that should *not* be annotated:**

俺	I ; (only) we
处处	everywhere
此	here
此间	here
大家	all ; everybody
当地	in the locality ; local
到处	everywhere
对方	opposite side ; the other party
多久	how long
个个	each one
各位	everyone
各自	each ; respective
何	who ; what
何时	what time ; when
后者	the latter
哪里	where
那里	there
你	you (singular)
你们	you (plural)
您	you (singular, formal)
其他	the others ; the rest
如此	such ; in this way
谁	who
什么 (麼)	what ; whatever
双边	both sides
双方	both sides ; the two parties
他人	another person ; others
我	I
我们	we
相互	mutual ; reciprocal
一方	one side
一切	all ; everything
以下	below ; following
有的	some
有些	some
咱们	we (inclusive)
这儿	here
这里	here
这样	such
自	self-
自我	self

Chapter 4

Hobbs

4.1 Introduction

The rule-based method chosen for resolving zero and overt pronominal anaphors in Chinese is the "Hobbs Algorithm".

In 1978 Jerry Hobbs proposed an algorithm for the resolution of pronominal coreference in English [20][1]. The performance of this algorithm has frequently been used as a reference point for computational methods of pronoun resolution in English ([14, 46, 48] and even in Spanish [13]).

The appeal of this algorithm is that it combines simplicity with a respectable performance. Behind the simplicity, however, are some non-trivial assumptions concerning the semantic knowledge provided by the system within which the algorithm runs in order to achieve the best performance.

In this work the algorithm will be run in three ways. The first way is the simplest, in which only syntactic knowledge (i.e., the parse tree) is used. The second incorporates limited gender and number information. The third attempts to draw on semantic information that is external to the surface syntax and to the operation

[1]The algorithm was actually published in a paper in 1976 [19], but this discussion refers to the algorithm and discussion from the more easily obtainable 1978 paper from *Lingua*.

of the algorithm itself in order to use selectional restrictions to filter the choices that the algorithm makes.

But why should an algorithm that was proposed for English work at all for Chinese? As will be illustrated in the next section, the algorithm exploits the fact that English is a subject-verb-object (SVO) language that relies more on word order than on morphology for determining the syntactic function of words and phrases. Chinese is also an SVO, word-order dependent language that has even less morphology than English.

The combination of the left-to-right, top-down operation of the algorithm and the SVO, fixed word order nature of the two languages makes it not unreasonable to try using the algorithm in Chinese to see how well it does.

Although Chinese and English are similar in being SVO languages, they differ in another significant parameter: Chinese is a pro-drop language, while standard English is not. Thus it also will be of interest to see how well the Hobbs algorithm performs when proposing antecedents for zero pronouns.

The rest of this chapter is organized as follows. First, we will describe two datasets on which the algorithm will be run, then a description of a smaller, "intersection" dataset will be given. Next the Hobbs algorithm and its execution will be explained, followed by a discussion of implementation issues. Finally, the results of executing the algorithm on the three datasets using the different sources of knowledge will be given and the results discussed.

4.2 The Data

4.2.1 The Penn Chinese Treebank

The Hobbs algorithm operates on parsed sentences. In order to evaluate its performance on zero pronouns as well as overt pronouns, it would be useful to have parses that already have the locations of the zero pronouns marked. Automatic parsers

Table 4.1: **Counts by syntactic level of coindexable third-person pronouns in the CTB**

Pronoun	M	M2	S	Total
he	315	31	138	484
he-they	32	16	74	122
she	20	9	38	67
she-they	0	0	4	4
it	14	22	50	86
it-they	7	3	29	39
Totals	**388**	**81**	**333**	**802**
% of total	**48.4%**	**10.1%**	**41.5%**	

do not yet produce empty categories in their outputs. The Penn Chinese Treebank (CTB) does, however, have overt strings that denote the positions of dropped arguments. Parses from Version 3 of the CTB were therefore used for experiments in resolving both overt third-person pronouns and zero pronouns. The CTB parses were annotated for coreference according to the guidelines discussed in Chapter 3[2].

In all there were 840 texts, with approximately 9500 sentences, containing 802 coindexed third-person pronouns and 1453 coindexed ***pro***.

4.2.2 2004 Chinese ACE Data

While using the gold-standard hand parses of the CTB allows us to focus on the problem of resolving zero pronouns without having to first find them, texts in "real world" applications are not so neatly pre-processed. Chinese texts first must be segmented into words (since written Chinese does not have spaces between words), then into sentences, then part-of-speech tagged, and finally parsed. Each of these steps when done automatically is a potential source of errors. More importantly for the work here, automatic parsers do not as a rule produce parses that have the

[2]The guidelines are on line: **http://www.cis.upenn.edu/~spc/guide.html**

dropped arguments (or any other empty category) explicitly denoted. Thus a step to identify the locations of the dropped arguments would be required. The machine parses still may be used to perform automatic pronominal coreference resolution for the overt pronouns, however.

The data obtained as part of Penn's participation in the 2004 Chinese ACE (CACE) bakeoff[3] were used as an approximation of a "real world" application. While the ACE annotation guidelines were quite different in many ways from the CTB annotation guidelines (see Section 3.8), the annotation of third-person pronouns is for the most part the same. Differences will be discussed in Section 4.2.3 below.

There are two parts to the CACE data, the training set and the final evaluation set. All development was done using the training set, with approximately one tenth of it used as the development set[4] for machine learning (to be discussed in Chapter 5). The 106 texts in the CACE training set that are taken from the CTB were kept separate. Each text in the CACE data had two parses available, one that was produced by the Bikel parser[5] and one produced by Bert Xue's maximum entropy parser[6].

Apart from the motivation of using the Hobbs algorithm to see how it would do on more realistic, machine-parsed data, there was a second purpose behind choosing to use the CACE data. That goal was to have enough data to train a supervised machine-learning model on the task.

The 2004 Chinese ACE training set (excluding the 106 CTB texts) has 1189 annotated overt third person pronouns, 74 of which are tagged generic or underspecified, in approximately 5356 sentences, in 540 documents. It also contains texts with a slightly wider range of sources, expanding upon the Xinhua News Agency articles of the CTB.

[3]See **http://www.nist.gov/speech/tests/ace/index.htm**
[4]aka "heldout" data
[5]See: **http://www.cis.upenn.edu/~dbikel/software.html**.
[6]Discussed in [51], see **http://verbs.colorado.edu/~xuen**.

Table 4.2: **Counts by syntactic level of third-person pronouns in the Bikel parses of CACE**

Pronoun	M	M2	S	Total
he	98	41	189	328
he-they	14	8	77	99
she	5	2	40	47
she-they	0	1	5	6
it	4	2	17	23
it-they	0	0	0	0
Totals	**121**	**54**	**328**	**503**
% of total	**24.1%**	**10.7%**	**65.2%**	

Table 4.3: **Counts by syntactic level of third-person pronouns in Xue parses of CACE**

Pronoun	M	M2	S	Total
he	94	41	193	328
he-they	10	12	77	99
she	5	1	41	47
she-they	0	1	5	6
it	2	5	16	23
it-they	0	0	0	0
Totals	**111**	**60**	**332**	**503**
% of total	**22.1%**	**11.9%**	**66.0%**	

The CACE evaluation set consists of 246 documents, with 2155 sentences, and 528 annotated third-person pronouns, 25 of which were classified as generic or underspecified. It was only used for the final evaluations to test the models that were developed using the training and heldout data.

4.2.3 Common Subset

In an attempt to compare just the effect of running on automatic parses *vs.* on hand parses on the performance of the algorithm, a third set of files was analyzed separately.

The 2004 Chinese ACE training data contain 106 texts that were taken from the first 100K of the Penn CTB. These were annotated for the ACE tasks according to the 2004 ACE guidelines[7].

Of these 106 files, nine had no overt pronouns (and did not even have ***pro*** in the CTB gold-standard parses) and were ignored. Two additional files of the 106 were ignored because the preprocessing steps in the Chinese ACE system had failed with the result that in each case both parses were useless for pronominal coreference.

One of these two files (**chtb_174**) had no ***pro*** in the gold CTB parse and just a single 他/ta1/*he* in the last sentence (the bad sentence). The other file (**chtb_254**) had only one overt demonstrative pronoun (此/ci3/*this*) that was tagged as discourse deictic in the CTB annotation. The gold key for the corresponding ACE file had no pronominals annotated.

After throwing out these pronounless and bad-parse files, 95 texts were left in the CTB-CACE intersection for comparing the execution of the Hobbs algorithm on the CTB gold-standard hand parses and on the automatic parses.

A special version of the Bikel parser was trained on data that explicitly excluded all 106 CTB texts used in the Chinese ACE training set. The Xue parser had been

[7]Found under the heading *ACE3* at:
http://projects.ldc.upenn.edu/ace/annotation/previous/

developed using n-fold cross validation, and was not retrained.

4.3 The Hobbs Algorithm

The "Hobbs Algorithm" "traverses the surface parse tree in a particular order looking for a noun phrase *of the correct gender and number*" ([20] page 315 ; emphasis added).

The steps are as follows:

1 begin at the NP node immediately dominating the pronoun

2 go up the tree to the first NP or S node encountered
  call this node **X**
  call the path to reach **X p**

3 traverse all branches below node **X** to the left of path **p**
  in a left-to-right, breadth-first manner
  propose as the antecedent any NP node that is encountered
  that has an NP or S node between it and **X**

4 if node **X** is the highest S node in the sentence,
  traverse the surface parse trees of previous sentences in the text
  in order of recency, starting with the most recent,
  left-to-right, breadth-first
  propose as antecedent the first NP encountered
  else continue with step (5)

5 from node **X** go up the tree to the first NP or S node encountered
  call this new node **X**, and
  call the path traversed to reach it from the original **X p**

6 if **X** is an NP node
  AND
  if the path **p** to **X** did not pass through the N-bar node that **X** immediately dominates,
  propose **X** as the antecedent

7 traverse all branches below node **X** to the *left* of path **p**

in a left-to-right, breadth-first manner

propose any NP node encountered as the antecedent

8 if **X** is an S node

traverse all branches of node **X** to the *right* of path **p**

in a left-to-right, breadth-first manner,

but do not go below any NP or S encountered

propose any NP node encountered as the antecedent

9 goto step 4

– – – – – – – – – – – – – –

While Hobbs calls the algorithm "naive", in that the steps proceed merely according to the structure of the parse tree, there are two meta-level points to consider in the execution of the steps. First, the algorithm counts on number and gender agreement. Second, in his paper, Hobbs proposes applying "simple selectional constraints" to the antecedents that the algorithm proposes, and illustrates their effectiveness in the sentence he uses to explain the operation of the algorithm:

> "The castle in Camelot remained the residence of the king until 536 when he moved it to London."

The parse tree Hobbs gives in his paper is shown in tree on page 73, and the steps that the algorithm takes are listed in Table 4.4.

When trying to resolve the pronoun "it" in this sentence, the algorithm would first propose "536" as the antecedent. But dates cannot move, so on selectional grounds it is ruled out. The algorithm continues and next proposes "the castle" as the antecedent. But castles cannot move any more than dates can, so selectional restrictions rule that choice out as well. Finally, "the residence" is proposed, and does not fail the selectional constraints.

In a similar fashion, resolving the pronoun "he" will follow the same steps (from step 2) that were followed for "it", but "the residence" will be rejected on gender grounds and the algorithm will settle on NP_6, "the king".

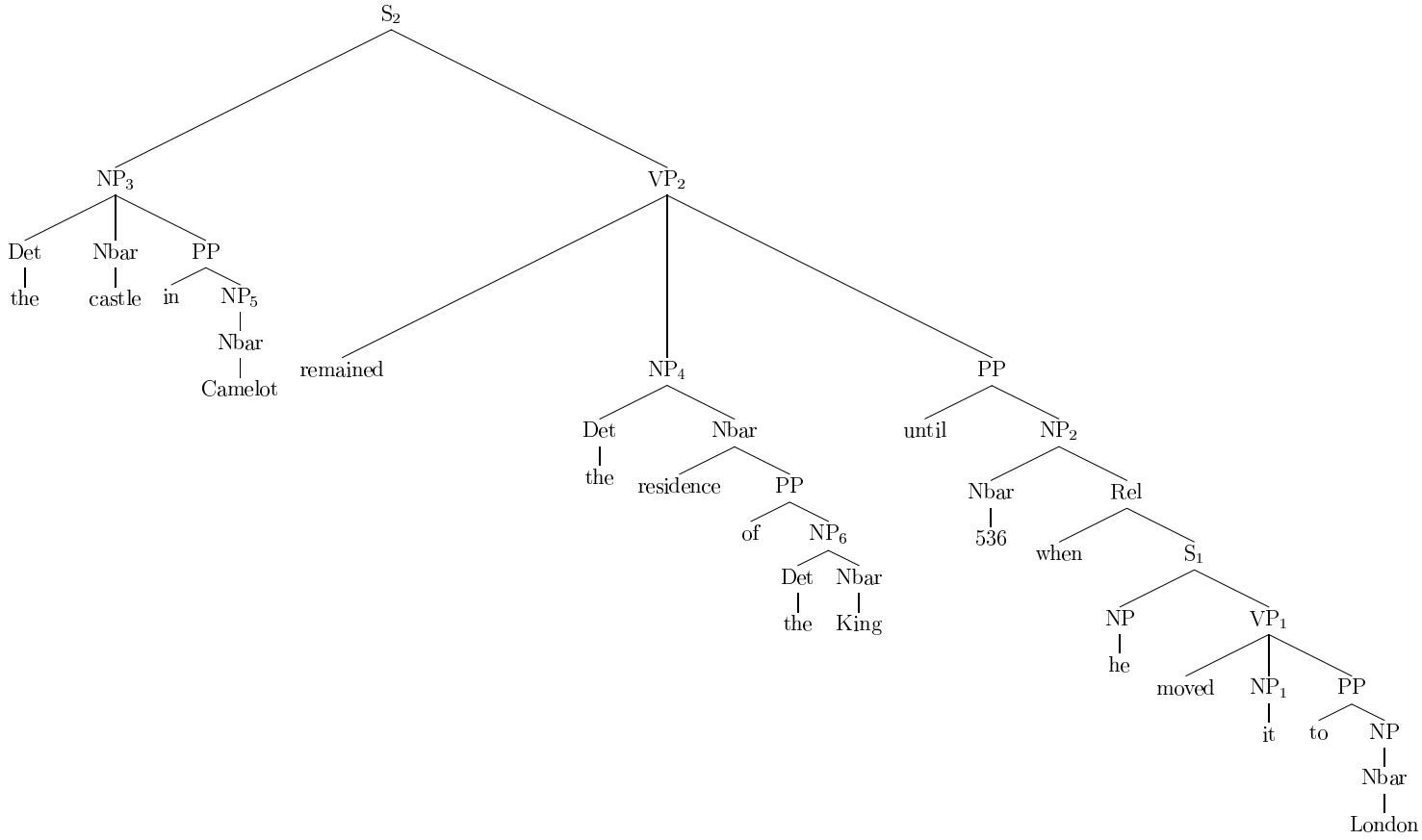
S2
NP3
VP2
Det
Nbar
PP
the
castle
in
NP5
Nbar
Camelot
remained
NP4
PP
Det
Nbar
until
NP2
the
residence
PP
Nbar
Rel
of
NP6
536
when
S1
Det
Nbar
NP
VP1
the
King
he
moved
NP1
PP
it
to
NP
Nbar
London

Table 4.4: **Steps taken by the Hobbs algorithm in resolving the "it" in the parse tree on page 73.**

step 1	begin at NP_1
step 2	go up to S_1 and mark it as **X**
step 3	search below **X**(S_1) to left of **p**, propose any NP that has NP or S between it and **X**(S_1)**:** there is *no* intervening NP, so continue
step 4	**X**(S_1) is not highest S-node in the sentence, so
step 5	go up to NP_2 and mark it as **X**
step 6	**X**(NP_2) *is* an NP node **and** the path **p** to **X**(NP_2) did *not* pass through the N-bar that **X**(NP_2) immediately dominates, so propose NP_2 ("536") as the antecedent **But** apply selectional constraints: (a) dates cannot move (b) places cannot move (c) large, fixed objects cannot move so reject NP_2 on constraint (a)
step 7	there is nothing to left of path **p** below **X**(NP_2)
step 8	**X**(NP_2) is not an S node
step 9	goto step 4
step 4	**X**(NP_2) is not an S node, much less the highest S node
step 5	go up to S_2 and mark it as **X**
step 6	**X**(S_2) is not an NP node
step 7	traverse the tree below **X**(S_2) to *left* of path **p**, breadth-first, left-to-right**:** propose NP_3 ("the castle"), but reject on selectional constraint (c), and continue breadth-first to NP_4 ("the residence of the king"), which is is not rejected by constraints

There is something noteworthy here with respect to the algorithm's tree traversal and the definition of "to the left of the path **p**". Note that node NP_4 is to the left of NP_2 but is a child of path node VP_2, which is an ancestor of NP_2, the first '**X**'. Thus the definition of "path" must be taken literally to be just the sequence of arcs between the nodes. That is, a node in the path that was reached from its right child and marked should *not* prevent its left child from being checked in the left-to-right, breadth-first search beginning at a higher node. The algorithm has been implemented to take this into account.

In addition, these "simple" selectional constraints require a fair amount of encoded world knowledge. The constraints invoked in the example presume a rather richly annotated lexical database and some kind of feature matching between predicates and their arguments. Dates and castles must be marked with the characteristics that they are fixed in time/space (after the digits 536 have first been recognized as a date in the first place), and the verb "move" must have the selectional restriction on its subject argument that the subject is not fixed in time or space.

In his paper, Hobbs reported the results of testing the algorithm on the pronouns "he", "she", "it" and "they", 300 instances in total (100 consecutive pronouns each from three different genres). He excluded "it" in time and weather constructions, as well as discourse deictic "it".

He found that the algorithm alone (*with* gender and number agreement) worked in 88.3% of the cases, and that the algorithm plus selectional restrictions correctly resolved 91.7% of the cases. But of the 300 examples, only 132 actually had more than one "plausible" antecedent nearby. When he tested the algorithm on just those 132 cases, 96 were resolved by the "naive" algorithm alone, a success rate of 72.7 percent. When selectional restrictions were added the algorithm correctly resolved 12 more, to give 81.8 percent.

Table 4.5: **Summary of Hobbs Paper Results**

Pronoun count	syntax only	with selectional constraints
300	88.3%	91.7%
132	72.7%	81.8%

4.4 Applying the Algorithm to Chinese

4.4.1 The CTB implementation

The Hobbs algorithm was first implemented to execute on the parses of the CTB[8]. The **S** label in the CTB is **IP**, so the two "markable" nodes from the point of view of the algorithm are **IP** and **NP** (relative clauses are labeled with **CP** in place of **Rel**).

The "basic" implementation simply follows the steps of the algorithm and proposes the first NP that meets the structural constraints named in the steps. Two corpus-specific restrictions were placed on the type of NP that would be allowed as an antecedent, however. NPs with either of the functional tags **-TMP** (temporal) or **-ADV** (adverbial) were prevented from being chosen as antecedents.

In the CTB, **NP-ADV** is most often the expression 其中/qi2zhong1, literally "of these-in the middle", which the CTB treats as one word meaning "among these". It never is the antecedent of a pronoun, but is frequently the initial NP of a clause. As for **NP-TMP**, time expressions can, of course, be antecedents of pronouns. In this corpus pronominal references to time expressions are extremely rare, however, and the number of misses of temporal antecedents is outweighed by the reduction in false positives due to clause- or sentence-modifying initial temporal NPs. In addition to NPs that were explicitly tagged **NP-TMP**, if the head noun had the POS tag NT (time noun), that NP was rejected. This corpus-specific filter could easily be put under run-time control.

[8]The bracketing guidelines [52] may be found at **http://www.cis.upenn.edu/~chinese/ctb.html**

This implementation was designed to be able to take into account, under run-time control, two sources of information beyond the parse structure alone. The first was a rough approximation of gender and number, and the second was a pair of databases with some semantic information.

Gender and number agreement

The first addition to the basic implementation is an approximation of number and gender agreement. The term "approximation" is used because Chinese nouns are not marked for number (with one very limited exception) or for gender. While the modern *written* versions of Chinese third-person pronouns do have number and gender[9], there is no morphology on verbs for either.

To determine gender, there are some gender-specific honorifics, as well as kinship terms that may be used. A table of these terms that we created for the Chinese ACE project was used. Unfortunately, both honorifics and kinship terms are scarce in the CTB. Although traditional Chinese names had some bias for what characters were used in names of girls and boys, modern names do not.

As for number agreement, there is an optional, plural "suffix" (们/men) that can be added to some nouns that can denote groups of people (e.g., 学生们/xue2sheng1men /*students* or 专家们/zhuan1jia1men/*experts*). In addition, one can sometimes glean number information from determiner phrases modifying head nouns.

A simple "parser" of determiner phrases was added to extract number information. It uses only part-of-speech information with some heuristics to determine number. It is conservative in proposing a definitive singular or plural reading (to override the default "agnostic" value assigned to a noun's number feature) in order to reduce false positives. To improve the coverage and accuracy would require using the actual parse structure to differentiate numbers and quantifiers that apply to the head noun itself from those that just belong to a modifier of the head noun[10].

[9] In spoken Chinese, only the number distinction is preserved.

[10] Chinese noun phrases are head final, and noun-noun compounding is common. A "ten-person

Because the number of cases that would be corrected from an incorrect to a correct antecedent as the result of a more robust implementation of number discrimination was expected to be small, this was left for future work.

Zero pronouns, of course, provide no clues about gender or number.

Selectional/semantic restrictions

The second source of information that can be utilized at run time consists of semantic features for nouns and the expected semantic features for the subjects of verbs.

The semantic features for nouns were obtained as follows. For her work on verb sense discrimination, Jinying Chen [4] created two semantic databases based on Hownet[11], the Rocling dictionary[12], and WordNet [11]. Each of the two supplied both a fine-grained and a coarse-grained classification for each noun (e.g., tool/artifact, language/communication). These databases were combined into a single noun-category database using just the coarse-grained set of categories. Each category was then mapped to a set of three-valued features. The set of features is: *human, animate, abstract/concrete, location*, and *organization*. Each feature may have the value true, false, or agnostic.

Feature assignments were made conservatively. For example, the category "event" is assigned the features: *human*:false, *animate*:false, *abstract*:true, *location*:agnostic, *organization*:agnostic. The two agnostic values avoid failures due to collisions of names of events that might overlap with the organizations that are sponsoring them or events that share the names of places where they are held.

A second database was built using the noun-category database, as follows:

1. get the head noun of each NP that is the subject of a verb (or multiple verbs) in each parsed sentence, look it up in the noun-category database, and write

committee" is the sort of phrase that illustrates the inadequacy of POS tags alone to determine number.

[11] **http://www.keenage.com**

[12] Refer to: **http://www.aclclp.org.tw**

out the subject-category:verb pair(s)

2. get the counts for each unique pair

3. construct a map for each verb with the following information:

```
<verb>
   <subject-noun-category> = <count>
   <subject-noun-category> = <count>
    ...
   <total> = <count>
```

These two databases, the noun-category feature database (henceforth just nouns database) and the verb-subject-categories database (hencefore just verbs database) were used by a constraint-checking filter.

When constraint checking was enabled at run time, the head noun of the proposed antecedent was looked up in the nouns database. For overt pronouns, if the noun's features had a conflict with the pronoun's features, it would be rejected. For example, the word "document(s)" falls under the category *communication*, which has the feature-value pairs *human*:false and *animate*:false. Both of those would conflict with the *human*:true and *animate*:true features of the male and female third person pronouns, so the constraint check would fail on whichever of those features was tested first whenever "document" was proposed as an antecedent for "she"/"he" (or one of their plural forms).

For null pronouns, if the ***pro*** was in the subject position (which it almost always is), then the subject categories for the verb were looked up in the verbs database. If the proposed antecedent's noun features had a conflict with the features for the verb's subject categories, then the antecedent was rejected.

Other implementation questions

Despite the SVO similarity between Chinese and English, we were interested to see if there might be structural differences between the languages that might require adapting its steps to Chinese. The most obvious place to look was in the placement of modifiers relative to the head noun in an NP. Although unplanned, it turned out that the annotation guidelines policy of annotating complex NPs at the parent level rather than at the head noun level (Section 3.6.1) actually made this a moot point because of the top-down nature of the tree traversal. That is, because the algorithm proposes an NP that contains both the modifier and the head, the difference between English and Chinese in head-modifier word order does not matter. The language specificity is pushed into the noun phrase head-finding function.

Step 6 of the algorithm relies on the syntactic formalism of a separate N-bar node to make the distinction between a PP that is the argument of a noun and a PP that is modifying the noun. The CTB does not use the N-bar notation, and an approximation was used for this step. The step "fired" only once when the algorithm was run on all 840 files of CTB-3, so this had a very minor impact.

Structurally, there are many sentences in the CTB that consist of just a sequence of parallel independent clauses, usually separated by commas or semicolons. That is, the top-level structure is

```
(IP
   (IP (NP)(VP))
   (PU ;)
   (IP (NP)(VP))
   ...
   (IP (NP)(VP))
   (PU .))
```

These multiple-clause sentences were not treated as multiple sentences, but were

left as single sentences.

A pronoun that was a subject at the top level of one of these **IP**s was assigned the syntactic level **M2** when results were analyzed. Such a pronoun is not a single matrix subject of a sentence, but on the other hand it is not really a subject in an embedded clause either.

4.4.2 The CACE implementation

In order to make it possible to run the Hobbs algorithm as part of the Penn Chinese ACE system, the algorithm was re-implemented in Java under the module of the CACE system that performed coreference resolution of named, nominal, and pronominal mentions. In addition, the coreference module's architecture was modified to make the pronominal coreference task independent of the general CACE system's coreference task (so that different resolvers could be plugged in), and to make it possible to score just the third-person pronouns[13].

The ACE annotation uses a stand-off annotation scheme. Each document's answer key contains a list of mentions (i.e., expressions in the text), that refer to the entities of interest in the document. The listed mentions are found in the document by their file offsets and extents, and each one is assigned the entity ID of the entity to which it refers. Mentions are categorized as *named*, *nominal*, or *pronominal* mentions, and classified as *specific*, *generic*, or *underspecified*. Another important bit of information associated with each mention is its ACE type. The ACE type is one of: **PER**SON, **ORG**ANIZATION, **GPE** (geopolitical entity), **LOC**ATION, **FAC**ILITY, **VEH**ICLE, or **WEA**PON. Only expressions that refer to an entity of one of these types are put into the answer key's list of mentions.

[13]The goal in ACE is to link together all mentions – named, nominal, and pronominal – for a single entity together into one chain, and that is the task that the original coreference resolution module peformed in the Penn system. The complex scoring for the system not only takes into account inaccurate links for a given entity (both false positives and false negatives), but also weights the mentions by mention type and ACE type. (Refer to *The ACE 2004 Evaluation Plan*, available at: **http://www.nist.gov/speech/tests/ace/ace04/index.htm**) It was therefore desireable to make a separate scoring mechanism to evaluate just the pronoun resolution function.

The program used the pronominal mentions for third-person pronouns from the answer key to determine what pronouns to resolve. It used machine parses of the documents for running the Hobbs algorithm. This meant that it was possible for the algorithm to fail almost immediately when a pronoun was incorrectly POS tagged and parsed, for example, as a verb or as punctuation, since there would be no parent NP.

To determine the correctness of the antecedent proposed by the algorithm, a match was sought between the antecedent parse NP constituent and the extent of a mention in the answer key. If a mention was found, the entity ID numbers were used for evaluating correctness. If the entity IDs differed or if no mention matched the antecedent NP, the antecedent was scored as incorrect.

Extra-syntactic information

As was the case for the CTB implementation of the algorithm, the program in CACE could be run with different controls to take into account different sources of information. In the "basic" mode, no agreement features or semantic databases were processed. In addition, there was no use of the ACE type information. The automatic parses did not have functional tags, so the only constraint beyond the structural ones of the algorithm itself on the NPs that could be picked as antecedents was a filter to exclude NPs that had noun heads with NT POS tags.

The first level of additional information consisted of the same approximations of number and gender that were used in the CTB implementation.

The second source of information was the ACE type. Any proposed antecedent had to be of the same ACE type as the pronoun. The ACE Type was used in lieu of semantic features for the Hobbs algorithm. Note that the ACE Type is a "perfect" source, in that it is hand annotated. It effectively does sense disambiguation of the mention. This is in contrast to the very noisy nouns and verbs databases that were used in the CTB.

Table 4.6: **Recency Baseline: CTB**

	Correct			**Wrong**				
	M	**M2**	**S**	**M**	**M2**	**S**	**Part.**	**Total**
he	221	9	36	94	22	102	-	484
he-they	15	4	14	17	12	59	1	122
she	13	1	12	7	8	26	-	67
she-they	0	0	1	0	0	3	0	4
it	10	6	8	4	16	42	-	86
it-they	1	2	5	6	1	24	0	39
Total overt	260	22	76	128	59	256	1	802
	67.0%	27.2%	22.9%	33.0%	72.8%	77.1%		
Count part.			359			443		802
as correct			**44.8%**			55.2%		

Part. means the algorithm picked one member of a plural entity
M: matrix-level
M2: subjects of independent clauses
S: pronouns in subordinate clauses (of any kind)

4.5 Baseline for Coreference Resolution Tests

A simple recency-based coreference resolver was used to provide a baseline for both the rule-based Hobbs implementations and the statistical resolvers that will be discussed in Chapter 5.

The baseline resolver chose the antecedent for each pronoun as follows. If the pronoun was the matrix subject of a sentence, it selected the first NP in the previous sentence, otherwise it picked the most recent NP, subject to binding constraints.

The results for the CTB gold-standard parses are given in Table 4.6. The column labeled **Part.** contains counts for plural pronouns that had one member of a split antecedent correctly identified. These were counted as correct in the total score only.

The columns labeled **M** represent anaphors at the matrix level, **M2** are subjects of parallel, independent clauses in multi-clause sentences (see page 80), and **S** stands for pronouns that were found in any kind of subordinate construction.

Table 4.7: **Recency Baseline: CTB – *pro* only**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
pro	83	96	314	102	242	563	1400
pro-plural	1	3	17	2	6	24	53
Total	84	99	331	104	248	587	1453
	44.7%	28.5%	36.1%				
			514			939	1453
			35.4%			64.6%	

The results for the gold parses indicate that recency of mention does not serve as a very good heuristic for non-matrix pronouns. The results from the machine parses show that with less than perfect parses, even selecting the previous subject does not do that well for matrix-level pronouns.

The results for the CACE parses are given in Tables 4.8 and 4.9.

The markedly worse performance of the recency baseline in the overall score for the machine parses relative to the gold-standard parses, 25.0% *vs.* 44.8%, may be in part attributed to a different distribution of syntactic levels in the two corpora as they are parsed. In the CTB, 48.4% of the pronouns are matrix subjects (Table 4.2.1), while the Bikel parses for CACE texts contained only 24.1% matrix-subject pronouns (Table 4.2) and the Xue parses just 22.1% for the same files (Table 4.3).

4.6 Tests of the Hobbs Algorithm

Although more than third-person pronouns were annotated in the CTB, we restricted these experiments to third-person pronouns and null pronouns because Hobbs himself only proposed the algorithm for third-person pronouns.

Table 4.8: **Recency Baseline: CACE Eval. Data – Bikel parses**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	31	4	54	67	37	135	328
he-they	2	2	11	12	6	66	99
she	3	1	10	2	1	30	47
she-they	0	1	1	0	0	4	6
it	1	1	4	3	1	13	23
it-they[1]	0	0	0	0	0	0	0
Total	37	9	80	84	45	248	503
	30.6%	16.7%	24.4%	69.4%	83.3%	75.6%	
			126			377	503
			25.0%			75.0%	

[1]Because there are no non-generic plural impersonal pronouns, this row will be omitted in the rest of the tables for the CACE pronouns.

Table 4.9: **Recency Baseline: CACE Eval. Data – Xue parses**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	35	8	49	59	33	144	328
he-they	1	3	13	9	9	64	99
she	2	0	7	3	1	34	47
she-they	0	1	1	0	0	4	6
it	1	2	3	1	3	13	23
Total	39	14	73	72	46	259	503
	35.1%	23.3%	22.0%	64.9%	76.7%	78.0%	
			126			377	503
			25.0%			75.0%	

4.6.1 Significance testing

The test that will be used here for determining statistical significance is McNemar's test. This is a non-parametric test suitable for analyzing matched pairs with binary labels[14]

For example, given the Table 4.10, one experiment assigns the labels **yes** and **no**. to a set of examples. Trial 2 labels the same set of examples. The pairs can be placed into one of the cells of the table depending on the labels assigned by the two trials.

Table 4.10: **Comparison of the outcomes of two trials**

		Trial 1	
		yes	**no**
Trial 2	**yes**	a	b
	no	c	d

Cells **a** and **d**, in which the two trials agree on the labeling, are of no interest. The null hypothesis of the test is that the counts of cells **b** and **c** are equal, i.e., the count of **yes-no** pairs is equal to the count of **no-yes** pairs.

The test calculates a P value using:

$$P = \frac{(|b-c|-1)^2}{(b+c)}$$

P has an asymptotic chi-square distribution with 1 degree of freedom.

4.6.2 Basic implementation run on gold-standard CTB parses

Tables 4.11 and 4.12 give the results of running the syntax-only version of the algorithm for coindexed third-person pronouns and for all coindexed zero pronouns in

[14] A nice brief summary of the test is given at:
http://www.fon.hum.uva.nl/Service/Statistics/McNemars_test.html
The on-line calculator that was used for the tests performed here (along with an explanation of the test and the formula for the calculation applied) may be found at:
http://www.graphpad.com/quickcalcs/McNemar1.cfm

Table 4.11: **Syntax only: CTB – overt pronouns**

	Correct			**Wrong**				
	M	**M2**	**S**	**M**	**M2**	**S**	**Part.**	**Total**
he	232	20	64	83	11	74	-	484
he-they	16	5	25	16	10	49	1	122
she	12	7	19	8	2	19	-	67
she-they	0	0	0	0	0	4	0	4
it	10	16	17	4	6	33	-	86
it-they	2	0	12	5	3	17	0	39
Total	272 70.1%	48 60.0%	137 41.1%	116 29.9%	32 40.0%	196 58.9%	1	802
Count part. as correct			458 **57.1%**			344 42.9%		802

the 840 files of CTB-3, respectively.

While the overall performance of 57.1% correct is certainly an improvement over the 44.8% of the recency score, the improvement shows up primarily in the **M2**- and **S**-level pronouns, since the recency strategy for matrix pronouns essentially mimics the behavior of the algorithm.

For the zero pronouns, the recency score for the matrix-level (44.7%) is actually better than in the Hobbs syntax-only run (40.2%), but the **M2** and **S** levels improved to a greater degree than the loss in performance of the **M** level, so Hobbs scores better overall (43.0% compared to 35.4% from Tables 4.12 and 4.7). An earlier experiment [8] on a smaller set of files had shown more promising results for Hobbs on ***pro***, with 76.3% correct at the matrix level and 53.2% overall. The accuracy for subordinate-level ***pro*** was about the same, at 43.3 percent.

4.6.3 Basic implementation run on CACE machine parses

Tables 4.13 and 4.14 show the results from running the basic syntax-only version of the Hobbs algorithm on machine parses of all files in the 2004 Chinese ACE

Table 4.12: **Syntax only: CTB – *pro***

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
pro	74	142	377	112	196	502	1403
pro-plural	2	8	23	1	1	19	54
	76	150	400	113	197	521	1457
	40.2%	43.2%	43.4%				
			626			831	1457
			43.0%			57.0%	

Table 4.13: **Syntax only: CACE Eval. Data – Bikel parses**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	53	17	80	45	24	109	328
he-they	6	3	17	8	5	60	99
she	4	0	14	1	2	26	47
she-they	0	1	1	0	0	4	6
it	2	1	8	2	1	9	23
Total	65	22	120	56	32	208	
	53.7%	40.7%	36.6%	46.3%	59.3%	63.4%	
			207			296	503
			41.2%			58.8%	

evaluation set. Results are given for the two different parses, and for third person pronouns only.

While the accuracy of the algorithm is an improvement over the poor baseline scores, the success was only around 41% for both sets of parses, far below the 57.1% for the gold-standard parses.

As was the case for the CTB Hobbs implementation, the algorithm performed with better accuracy on the matrix level pronouns than it did on the subordinate or independent-clause level pronouns.

Table 4.14: **Syntax only: CACE Eval. Data – Xue parses**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	53	21	73	41	20	120	328
he-they	3	5	20	7	7	57	99
she	3	1	12	2	0	29	47
she-they	0	1	0	0	0	5	6
it	1	4	7	1	1	9	23
Total	60	32	112	51	28	220	503
	54.1%	53.3%	33.7%	45.9%	46.7%	66.3%	
			204			299	503
			40.6%			59.4%	

4.6.4 Basic implementation run on a subset common to CTB and CACE

While it is difficult to compare the performance of the algorithm when run on the CTB data *vs.* the CACE data in general, the 95 files that appear in both sets (discussed in Section 4.2.3 above) might provide some insight into the effect that using machine parses, instead of "gold-standard" parses, has on the success of the algorithm. Table 4.15 shows the distribution of the overt third-person pronouns from the 95 files that were coindexed with an antecedent in the CTB. Table 4.16 lists the pronouns after removing the ones that are not found in the CACE answer keys for the 95 texts.

Pronoun mismatches

Discrepancies in pronoun counts between the CACE annotated data and the CTB annotated data can arise in a number of ways. In the ACE tasks, only those mentions for entities that belonged to one of the seven target types were annotated. Thus, if a pronoun referred to an entity that was not one of the ACE types, it would not be annotated. Since PERSON and ORGANIZATION were two of the types, this meant

Table 4.15: **Counts for 3rd-person pronouns in common subset that were annotated in the CTB**

	M	M2	S	Total
he	46	5	25	76
he-they	3	3	4	10
she	1	0	0	1
she-they	0	0	0	0
it	5	4	10	19
it-they	1	0	2	3
	56	**12**	**41**	**109**

Table 4.16: **Counts for 3rd-person pronouns in common subset that were coded in CACE**

	M	M2	S	Total
he	46	5	25	76
he-they	3	3	4	10
she	1	0	0	1
she-they	0	0	0	0
it	5	4	5	14
it-they	1	0	2	3
	56	**12**	**36**	**104**
	53.9%	11.5%	34.6%	

Table 4.17: **Common subset: syntax only: CTB parses**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	38	4	10	8	1	15	76
he-they	2	1	2	1	2	2	10
she	1	0	0	0	0	0	1
she-they	0	0	0	0	0	0	0
it	4	3	2	1	1	3	14
it-they	1	0	0	0	0	2	3
Total	46	8	14	10	4	22	104
	82.1%	66.7%	38.9%	17.9%	33.3%	61.1%	
			68			36	104
			65.4%			34.6%	

that it was extremely unlikely that the gendered third person pronouns would not be annotated. The pronoun "it" or plural "it" might possibly refer to an entity that was not of an ACE type, however.

In ACE, mentions were classified as being *specific*, *generic*, or *underspecified*. Generic and underspecified pronouns were ignored by the system. It is possible that a pronoun classified as generic in ACE would have a category label in the CTB annotation as well (e.g., **#AMB**, **#INFR**, or **#ARB**). On the other hand, if the generic pronoun is a bound pronoun, in CTB both the pronoun and the expression naming its type would be coindexed.

It is also possible that the CTB annotator assigned a category label to a pronoun that the CACE annotator coindexed as part of an entity.

In addition, there was no requirement that a pronoun in ACE be coindexed with an expression that was not pronominal (i.e., a nominal or a name mention). Thus, in one file there are two instances of he-they that are coindexed as belonging to the same entity, but there is no other named or nominal expression that refers to that entity.

While the CACE "gold" key files were used for the list of pronouns to be resolved

Table 4.18: **Common subset: syntax only: CACE Bikel parses**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	30	4	9	14	1	18	76
he-they	2	1	2	0	4	1	10
she	1	0	0	0	0	0	1
she-they	0	0	0	0	0	0	0
it	2	4	3	1	1	3	14
it-they	1	0	0	0	1	1	3
Total	36 70.6%	9 56.2%	14 37.8%	15 29.4%	7 43.8%	23 62.2%	
			59 **56.7%**			45 43.3%	104

Table 4.19: **Hobbs: Syntax-only version: Common Subset**

Parse Source	**Counts and Percent Correct**			
	M	**M2**	**S**	**Total**
CTB parses	**82.1%** 46/56	**66.7%** 8/12	**38.9%** 14/36	**65.4%** 68/104
Bikel parses	**70.6%** 36/51	**56.2%** 9/16	**37.8%** 14/37	**56.7%** 59/104

by each method, the machine parses might not have arrived at the correct phrase structure surrounding the pronoun.

Finally, in any annotated data set there are omissions or mis-labelings simply due to human error[15].

Tables 4.17 and 4.18 give the detailed results of running the syntax-only versions of the algorithm on the intersection set, and Table 4.19 summarizes the scores.

To determine whether or not the difference in accuracy is significant, the two answers for each pronoun were compared, one from the CTB parse and one from the Bikel parse. Using McNemar's test the difference in overall performance had a P

[15]and comparing different annotations of the same sources can often be very useful in debugging!

value of 0.0055, which is very significant. The difference for matrix level pronouns is also significant (P = 0.0233), but the **M2** and **S** levels are not significantly different.

This small test suggests that the quality of the parse does make a significant difference where structure contributes the most, for the matrix-level pronouns.

4.6.5 Syntax plus agreement

In spite of their limited nature, adding the simple number and gender agreement checks to the algorithm did result in an overall improvement that was statistically significant, with P = 0.0094, in the CTB data, even though the difference in actual percentages is smll between 57.1% for the syntax only (Table 4.11) and 58.4% for syntax plus agreement (Table 4.20).

The improvement came primarily from an increase in accuracy at the subordinate level that was also significantly different from the syntax-only score (P = 0.0233).

In the noisier CACE parses, the agreement constraints increased overall performance only from 41.2% to 42.1% in the Bikel parses (Tables 4.13 to 4.21) and 40.6% to 41.4% in the Xue parses (Tables 4.14 and 4.22). Neither difference was significant. The improvement at the **S**-level alone for the Bikel parses was significant, however (P = 0.0455).

It is noteworthy that in all three sets of parses that most of the improvement came from the subordinate level pronouns.

4.6.6 Syntax plus semantics only

The "semantic constraints" that were used for the CTB parses were described on page 78. For the CACE parses, the semantic constraint was implemented as a check of the ACE types of the pronoun and its antecedent. If the antecedent had an ACE

Table 4.20: **Agreement only: CTB**

	Correct			**Wrong**				
	M	**M2**	**S**	**M**	**M2**	**S**	**Part.**	**Total**
he	233	19	67	82	12	71	-	484
he-they	17	5	25	15	10	49	1	122
she	12	7	21	8	2	17	-	67
she-they	0	0	0	0	0	4	0	4
it	11	17	19	3	5	31	-	86
it-they	2	0	12	5	3	17	0	39
Total All	275	48	144	113	32	189	1	802
	70.9%	60.0%	43.2%	29.1%	40.0%	56.8%		
Count part. as correct	468			334				802
	58.4%			41.6%				

Table 4.21: **Agreement only: CACE – Bikel Parses**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	52	16	81	46	25	108	328
he-they	6	3	18	8	5	59	99
she	4	0	19	1	2	21	47
she-they	0	1	1	0	0	4	6
it	2	1	8	2	1	9	23
Total	64	21	127	57	33	201	503
	52.9%	38.9%	38.7%	47.1%	61.1%	61.3%	
			212			291	503
			42.1%			57.9%	

Table 4.22: **Agreement only: CACE – Xue Parses**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	53	21	73	41	20	120	328
he-they	3	5	21	7	7	56	99
she	4	1	14	1	0	27	47
she-they	0	1	0	0	0	5	6
it	1	4	7	1	1	9	23
Total	61	32	115	50	28	217	503
	55.0%	53.3%	34.6%	45.0%	46.7%	65.4%	
			208			295	503
			41.4%			58.6%	

type, it had to match the pronoun's ACE type or it was rejected, and the algorithm continued until it found an antecedent that did not have a conflict.

While the total accuracy increased after adding semantic constraints to the CTB implementation, it did so only through an improvement in the performance on pronouns in subordinate clauses more than the decrease in the accuracy for matrix subject pronouns. Even though the overall increase in score was not significant, the improvement at the **S** level was very significant, with a P value of 0.0010. This is consistent with the hypothesis that pronouns at subordinate levels are subject more to the semantic context than to the syntactic structure, while the opposite is true for the matrix pronouns.

In contrast to these effects on the CTB parses, the constraint that the antecedent's ACE type match the pronoun's ACE type improved the accuracy for all levels of pronoun in the CACE parses, although in both cases more than half of the improvement was due to the **S**-level accuracy. That all levels improved is not surprising because the ACE type here is taken from the hand-coded answer key[16], whereas the semantic features used in the CTB nouns and verbs databases are automatically generated using coarse-grained categories.

For the zero ***pro***s, unfortunately this first experiment with selectional constraints on the potential antecedents as described on page 78 gained only a single, matrix-level pronoun (Table 4.24 compared to 4.12).

4.6.7 Syntax plus both agreement and semantics

The final experiment in this group was to test both gender and number agreement in combination with each implementation's semantic constraints. Combining agreement

[16]In the Penn CACE system, there were statistical named entity and nominal entity taggers that assigned the types. The gold keys were used here to avoid adding another source of error to these experiments.

Table 4.23: **Semantics only: CTB**

	Correct			**Wrong**				
	M	**M2**	**S**	**M**	**M2**	**S**	**Part.**	**Total**
he	218	17	67	97	14	71	-	484
he-they	19	10	42	13	5	30	3	122
she	11	4	20	9	5	18	-	67
she-they	0	0	3	0	0	1	0	4
it	10	16	18	4	6	32	-	86
it-they	2	0	11	5	3	18	0	39
Total	260	47	161	128	33	170	3	802
	67.0%	58.8%	48.6%	33.0%	41.2%	51.4%		
Count part. as correct			471			331		802
			58.7%			41.3%		

Table 4.24: **Semantics only: CTB – *pro***

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
pro	75	142	377	111	196	502	1403
pro-plural	2	8	23	1	1	19	54
	77	150	400	112	197	521	1457
	40.7%	43.2%	43.4%				
			627			830	1457
			43.0%			57.0%	

Table 4.25: **ACE Type match only: Bikel parses**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	58	18	84	40	23	105	328
he-they	6	5	20	8	3	57	99
she	4	0	16	1	2	24	47
she-they	0	1	1	0	0	4	6
it	2	1	9	2	1	8	23
Total	70	25	130	51	29	198	
	57.9%	46.3%	39.6%	42.1%	53.7%	60.4%	
			225			278	503
			44.7%			55.3%	

Table 4.26: **ACE Type match only: Xue parses**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	56	23	81	38	18	112	328
he-they	3	6	25	7	6	52	99
she	3	1	13	2	0	28	47
she-they	0	1	0	0	0	5	6
it	1	4	7	1	1	9	23
Total	63 56.8%	35 58.3%	126 38.0%	48 43.2%	25 41.7%	206 62.0%	503
			224 **44.5%**			279 55.5%	503

Table 4.27: **Agreement with Semantics: CTB**

	Correct			Wrong				
	M	**M2**	**S**	**M**	**M2**	**S**	**Part.**	**Total**
he	219	16	72	96	15	66	-	484
he-they	20	10	42	12	5	30	3	122
she	11	5	22	9	4	16	-	67
she-they	0	0	3	0	0	1	0	4
it	11	17	18	3	5	32	-	86
it-they	2	0	11	5	3	18	0	39
Total	263 67.8%	48 60.0%	168 50.8%	125 32.2%	32 40.0%	163 49.2%	3	802
Count part. as correct			482 **60.1%**			320 39.9%		802

and the use of the noun and verb databases did improve the overall performance of the CTB implementation from 57.1% (Table 4.11) to 60.1% (Table 4.27), a significant difference, with P=0.0486.

All of the improvement was due to the extremely significant improvement in **S**-level pronouns ($P < 0.0001$).

For the CACE parses, adding agreement to the ACE type constraint did improve overall accuracy in both sets of parses, but not significantly. Again, the **S**-level pronouns were where the improvement was observed (Tables 4.28 and 4.29).

Table 4.28: **ACE Type match plus Agreement: Bikel parses**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	58	17	84	40	24	105	328
he-they	6	5	20	8	3	57	99
she	4	0	21	1	2	19	47
she-they	0	1	1	0	0	4	6
it	2	1	9	2	1	8	23
Total	70	24	135	51	30	193	503
	57.9%	44.4%	41.2%	42.1%	55.6%	58.8%	
			229			274	503
			45.5%			54.5%	

Table 4.29: **ACE Type match plus Agreement: Xue parses**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	56	23	81	38	18	112	328
he-they	3	6	25	7	6	52	99
she	4	1	15	1	0	26	47
she-they	0	1	0	0	0	5	6
it	1	4	7	1	1	9	23
Total	64	35	128	47	25	204	503
	57.7%	58.3%	38.6%	42.3%	41.7%	61.4%	
			227			276	503
			45.1%			54.9%	

4.6.8 Discussion

Table 4.30 shows a summary of the results that have been presented in this chapter. The counts of the number of pronouns at each syntactic level are also given (number correct / the total number for that level).

The difference in performance of the algorithm by syntactic level and the differential effects of the two extra sources of information on accuracy clearly suggest that a one-method-fits-all, rule-based approach to anaphora resolution will not succeed.

These results are consistent with the observations made by Miltsakaki in her 2002 paper [33]. Taking a main clause and all its dependent clauses as a unit, she found that there were different mechanisms needed to account for (1) topic continuity from unit to unit (inter-sentential), and (2) focusing preferences within a unit (intra-sentential). Topic continuity was best modeled structurally (as the Hobbs algorithm does), but the semantics and pragmatics of verbs and connectives were more important than syntax *within* a unit.

In our results the anaphors that are subjects of matrix clauses tend to resolve inter-sententially (that is, Step 4 of the algorithm is the resolving condition), while the anaphors in subordinate constructions are more likely to have intra-sentential antecedents. That the strictly structural version of the Hobbs algorithm used here performed better for matrix-level anaphors (the best scores for **M**-level pronouns are 70.9% for the CTB (row G), 57.9% for Bikel parses (either row K or N), and 57.7% for Xue parses (row O)) than it did for anaphors in subordinate constructions (in which the best scores were 50.8% for CTB (row M), 38.7% and 41.2% for CACE (rows H and N)) agrees with Miltsakaki's findings.

The hypothesis that applying selectional restrictions to filter potential antecedents would improve the performance of the Hobbs algorithm to a greater degree with respect to the anaphors appearing in subordinate constructions than it would for those appearing at the matrix level was borne out by these experiments.

Adding limited gender and number agreement to the basic CTB implementation

increased accuracy for subordinate-level pronouns to a greater degree than it did for matrix pronoun resolution. That is comparing row (G) to row (D), the gain in both count and percentage points was better for **S** pronouns than for those at the matrix level **M** (and significant only in the **S** case).

Adding semantics alone showed the effect better. There was a 7.5% increase in **S**-level resolution in row (J) (48.6%) above the syntax-only results in row (D) (41.1%), while the accuracy on matrix pronouns *decreased* (70.1% to 67.0%), so that the overall improvement from adding just the semantics component (from 57.1% to 58.4%, not significant) is attributed to the improvement at the **S** level alone (from 41.1% to 48.6%, very significant, at P = 0.0010). When agreement and semantics were both added, as summarized in row (M), the agreement appears to have mitigated somewhat the damage that the semantic constraints did to the performance at the matrix level.

The **M2** level in the CTB data seems to be impervious to both agreement and semantics, holding steady around 60% correct.

For the machine parses in CACE, adding agreement checks to the syntactic rules improved the overall score slightly, but only the **S** increase in the Bikel parses was significant. In contrast, adding the selectional restriction that ACE types match, in lieu of using semantic databases, improved the accuracy of the algorithm at all three syntactic levels in both sets of parses.

In the next chapter we will examine a machine learning approach to pronoun resolution to see what the effects of different sources of feature information can have on the performance of the models.

Table 4.30: **Hobbs Algorithm Results Summary**

	Information Source	**Counts and Percent Correct**			
		M	**M2**	**S**	**Total**
A	Recency baseline	**67.0%**	**27.2%**	**22.9%**	**44.8%**
	CTB	260/388	22/81	76/332	359/802
B	Recency baseline	**30.6%**	**16.7%**	**24.4%**	**25.0%**
	CACE Bikel parses	37/121	9/54	80/328	126/503
C	Recency baseline	**35.1%**	**23.3%**	**22.0%**	**25.0%**
	CACE Xue parses	39/111	14/60	73/332	126/503
D	Syntax only	**70.1%**	**60.0%**	**41.1%**	**57.1%**
	Hobbs - CTB	272/388	48/80	137/333	458/802
E	Syntax only	**53.7%**	**40.7%**	**36.6%**	**41.2%**
	CACE Bikel	65/121	22/54	120/328	207/503
F	Syntax only	**54.1%**	**53.3%**	**33.6%**	**40.5%**
	CACE Xue	60/111	32/60	112/333	204/50
G	Syn + Agreement	**70.9%**	**60.0%**	**43.2%**	**58.4%**
	CTB	275/388	48/80	144/333	468/802
H	Syn + Agreement	**52.9%**	**38.9%**	**38.7%**	**42.1%**
	CACE Bikel	64/121	21/54	127/328	212/503
I	Syn + Agreement	**55.0%**	**53.3%**	**34.6%**	**41.4%**
	CACE Xue	61/111	32/60	115/333	208/503
J	Syn + Semantics	**67.0%**	**58.8%**	**48.6%**	**58.7%**
	CTB	260/388	47/80	161/331	471/802
K	Syn + ACE Type	**57.9%**	**46.3%**	**39.6%**	**44.7%**
	CACE Bikel	70/121	25/54	130/328	225/503
L	Syn + ACE Type	**56.8%**	**58.3%**	**38.0%**	**44.5%**
	CACE Xue	63/111	35/60	126/333	224/503
M	Syn + Sem + Agr	**67.8%**	**60.0%**	**50.8%**	**60.1%**
	CTB	263/388	48/80	168/331	482/802
N	Syn + ACE + Agr	**57.9%**	**44.4%**	**41.2%**	**44.5%**
	CACE Bikel	70/121	24/54	135/328	229/503
O	Syn + ACE + Agr	**57.7%**	**58.3%**	**38.6%**	**45.1%**
	CACE Xue	64/111	35/60	128/332	227/503

Chapter 5

Maximum Entropy Models

5.1 Introduction

Chapter 4 outlined the results from running a well known, rule-based algorithm, the Hobbs algorithm, to resolve both third-person and zero pronominal anaphors in Chinese texts.

The best performance on finding the antecedents for overt matrix pronouns was 70.9% (Table 4.30, row G) for the CTB syntax plus agreement version of the algorithm, and 57.9% (row N) and 57.7% (row O) for the Bikel and Xue parses, respectively, from the version with syntax plus agreement and the ACE type matching[1] constraint.

Accuracy for pronouns in subordinate constructions was worse. The best scores for the CACE data were also obtained when both agreement and ACE type were used as constraint filters, but the scores were still very low, at 41.2% for Bikel parses (row N) and 38.6% for Xue parses (row O). For the CTB implementation, the best result for the subordinate pronouns of 50.8% was also obtained using both agreement and semantic features (row M).

[1] or, more accurately, not conflicting

Machine learning methods have been used with good results for anaphora resolution in both English and Japanese (for example, see [1, 14, 36, 43]). We were interested to see how well a statistical machine learning approach would work for Chinese pronouns. In particular, we wanted to see if a model could be trained that performed as well on the subordinate level pronouns as it did on the matrix-level pronouns.

We began by using just basic syntactic and surface features. We then added agreement and semantic features to see how much they contributed to the performance of the models.

5.2 Implementation

In Section 4.4.2 we explained that we modified the architecture of the general coreference module of Penn's 2004 Chinese ACE system in order to experiment with different pronominal coreference resolvers. The modifications allowed control over not only what kind of pronoun resolver was used, but also enabled scoring of the performance of the pronoun coreference resolution component alone.

The original pronoun resolver component part of the Penn Chinese ACE coreference module was a maximum entropy model built using the Mallet package[2]. Because pronouns were discounted so much by the ACE scoring algorithm[3], and because the range of pronominal types was so great in the ACE guidelines (discussed in Section 3.8.3), the resolver component received almost no engineering effort, and actually made the overall system's score worse when it was used (by linking together mentions that should not be linked, thereby increasing the false positives, which were heavily penalized).

Nevertheless, the skeleton of that resolver provided a framework for bootstrapping

[2]Found at: **http://mallet.cs.umass.edu**

[3]The document *The ACE 2004 Evaluation Plan* may be found at: **http://www.nist.gov/speech/tests/ace/ace04/index.htm**

the models built here. It was modified to train on only third-person pronouns.

The general architecture of the system's coreference resolution module, carried over from Penn's 2003 Chinese ACE system, was one in which "entities" were built incrementally as mentions were encountered in the text. An "Entity" object consisted of a collection of one or more "Mention"s, each of which ideally referred to the same world entity. Thus the objects being compared in a machine-learning event were a mention and an entity, rather than a mention and another mention[4]. In this way, entities could collect in one place information about the real world object to which the different mentions referred. One mention might provide gender by way of an honorific, while another might provide number from a determiner phrase. This is the approach that Tom Morton used, with good results, in English [36].

In addition, a separate kind of model was developed that trained on events based on mention-to-mention comparisons, which is an approach frequently used in the literature. The motivation for creating another pronominal model type to "plug in" to the system's coreference module was to try to compare the two different approaches on the same set of data. Because this comparision was not the primary focus of this work it was not fully worked out, but the differences in performance between the two models do hint at the strengths of the entity-based architecture.

The two model types are referred to as entity-mention and mention-mention models in the following discussion.

5.3 Feature Sets for Different Models

Just as the Hobbs rule-based resolver was run using different sources of information, different maximum-entropy models were trained with features that represented different sources or levels of information. The first set of features represented purely structural and surface features, and the models trained with them were dubbed the

[4]This is not unlike a system that the IBM team described at the ACE 2004 meeting and that was published in [31]

"syntax only" models.

Using the structural features as a base, other small sets of features were added alone or in combination to see if adding features with more semantic content would improve the performance.

The first set that was added contained features derived from number and gender agreement, when available. Although not features in the model-training sense, the ACE type of potential antecedents was again used as a semantic constraint in selecting training and evaluation examples. A second set of model features, called the "semantic features" in this chapter, used the same kind of nouns database and verbs database that were described in Section 4.4.1 under **Selectional/semantic restrictions** (pages 78*ff*) to construct features based on noun-category matches. Finally, there were features that were mixed semantic and pseudo-pragmatic information, which are called "pragmatic" here just to distinguish them from the semantic features using the databases.

The following sections describe each group of features and give the results from testing the models trained on them.

5.3.1 Syntactic Features Only

The first models were built using just structural or surface features. The features included the following:

- the pronoun's head word and POS
- string matching between the pronoun and potential antecedent
- the clause depth and sentence depth of the pronoun
- the pronoun's sentence position: start/end of sentence, start/end of clause, or somewhere between

- a paired feature of the potential antecedent's sentence position and the pronoun's sentence position
- whether or not the pronoun and antecedent are in the same sentence
- whether or not the pronoun and the antecedent are at the start of consecutive sentences
- the distance in words between the pronoun and the antecedent
- the "distance" in mention counts between the pronoun and the antecedent in the list of mentions recognized up to that point in the parse
- whether or not the pronoun and antecedent are separated by only a copular verb (using POS "VC")
- whether or not the antecedent c-commands the pronoun

The accuracy of these first models was encouraging. As Tables 5.1, 5.2, 5.3, and 5.4 show, the overall performance for all four models was above 64% accuracy, with even higher performance for the **M2** level. This was already a good improvement over the Hobbs algorithm's 41% accuracy for the same data with just the syntactic information (Tables 4.13 and 4.14).

Contributing to these early results were the accuracies of the models on the subordinate-level pronouns. These ranged from 64.5% (Table 5.4) to 66.5% (Table 5.1).

Table 5.1: **Syntactic features only: Bikel entity-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	70	35	144	28	6	45	328
he-they	8	5	31	6	3	46	99
she	5	2	28	0	0	12	47
she-they	0	1	4	0	0	1	6
it	1	2	11	3	0	6	23
it-they[1]	0	0	0	0	0	0	0
Total	84	45	218	37	9	110	
	69.4%	83.3%	**66.5%**	30.6%	16.7%	33.5%	
			347			156	503
			69.0%			31.0%	

[1]**Note:** because there were no non-generic plural neutral pronouns, this row will be omitted from the remaining tables.

Table 5.2: **Syntactic features only: Bikel mention-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	63	30	139	35	11	50	328
he-they	5	4	27	9	4	50	99
she	2	1	31	3	1	9	47
she-they	0	1	3	0	0	2	6
it	3	2	14	1	0	3	23
Total	73	38	214	48	16	114	
	60.3%	70.4%	65.2%	39.7%	29.6%	34.8%	
			325			178	503
			64.6%			35.4%	

Table 5.3: **Syntactic features only: Xue entity-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	69	33	146	25	8	47	328
he-they	7	7	30	3	5	47	99
she	4	1	31	1	0	10	47
she-they	0	1	3	0	0	2	6
it	1	3	10	1	2	6	23
Total	81	45	220	30	15	112	
	73.0%	75.0%	66.3%	27.0%	25.0%	33.7%	
			346			157	503
			68.8%			31.2%	

Table 5.4: **Syntactic features only: Xue mention-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	58	34	139	36	7	54	328
he-they	1	6	28	9	6	49	99
she	2	1	32	3	0	9	47
she-they	0	1	3	0	0	2	6
it	1	5	12	1	0	4	23
Total	62	47	214	49	13	118	
	55.9%	78.3%	64.5%	44.1%	21.7%	35.5%	
			323			180	503
			64.2%			35.8%	

5.4 Syntactic Features Plus Agreement

The first level of information added to the structural features was agreement, namely number and gender, what little information there was. The same approaches for obtaining gender and number agreement were used as were described on page 77 above. There were five features for gender and five for number: both pronoun and antecedent were marked, match yes and no; one was marked but the other wasn't; and neither was marked.

The overall scores improved for the entity-mention models (69.0% to 70.4% for the Bikel parses (Tables 5.1 and 5.5) and 68.8% to 70.8% for Xue parses (Tables 5.3 and 5.7)), but not significantly. The improvement was primarily due to the **S**-level pronouns: in the case of the Bikel parses, that improvement outweighed a decrease in accuracy at the matrix level. The corresponding mention-mention models did not improve or got worse.

The fact that the entity-mention model benefited more (and consistently) from agreement features makes sense, and illustrates one of the strengths of this type of model, in that while one mention may not exhibit a characteristic of the entity, another might. By matching a new pronoun against the entire collected knowledge for an entity up to that point in the parse, the chances are greater for matching a characteristic such as gender or number.

5.5 Syntactic Features Plus ACE Type, With and Without Agreement Features

Rather than a feature *per se*, the first source of "semantic" information that was used was to limit the potential antecedent entities or mentions to those that had the

Table 5.5: **Syntactic and Agreement features: Bikel entity-mention**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	68	35	147	30	6	42	328
he-they	8	5	35	6	3	42	99
she	4	2	31	1	0	9	47
she-they	0	1	3	0	0	2	6
it	2	2	11	2	0	6	23
Total	82	45	227	39	9	101	
	67.8%	83.3%	69.2%	32.2%	16.7%	30.8%	
			354			149	503
			70.4%			29.6%	

Table 5.6: **Syntactic and Agreement features: Bikel mention-mention**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	60	28	134	38	13	55	328
he-they	6	5	34	8	3	43	99
she	3	2	28	2	0	12	47
she-they	0	1	2	0	0	3	6
it	2	2	14	2	0	3	23
Total	71	38	212	50	16	116	
	58.7%	70.4%	64.6%	41.3%	29.6%	35.4%	
			321			182	503
			63.8%			36.2%	

Table 5.7: **Syntactic and Agreement features: Xue entity-mention**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	71	32	145	23	9	48	328
he-they	7	7	36	3	5	41	99
she	4	1	34	1	0	7	47
she-they	0	1	4	0	0	1	6
it	1	3	10	1	2	6	23
Total	83	44	229	28	16	103	
	74.8%	73.3%	69.0%	25.2%	26.7%	31.0%	
			356			147	503
			70.8%			29.2%	

Table 5.8: **Syntactic and Agreement features: Xue mention-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	59	26	138	35	15	55	328
he-they	2	8	36	8	4	41	99
she	3	1	27	2	0	14	47
she-they	0	1	4	0	0	1	6
it	1	3	14	1	2	2	23
Total	65	39	219	46	21	113	
	58.6%	65.0%	66.0%	41.4%	35.0%	34.0%	
			323			180	503
			64.2%			35.8%	

same ACE type as the pronoun. This constraint alone, when added to the model with just syntactic features, improved the performance of both kinds of model over the syntax-only models to a very significant degree.

Both entity and mention models for the Bikel parses scored at or above 76.9% when the ACE type constraint alone was added (Tables 5.9 and 5.10), compared to the best scores of 69.0% for syntax only (Table 5.1) and 70.4% for syntax plus agreement (Table 5.5) from the entity-mention model.

The models built on the Xue parses scored in the same range, with 79.5% for the entity model (Table 5.11) and 77.7% for the mention-mention model (Table 5.12), compared to 70.8% (Table 5.7) and 64.5% (Table 5.8) for syntax plus agreement.

For all four models, the improvement was seen across all three syntactic levels but was most significant for the subordinate-level pronouns.

Keeping the ACE type constraint and adding the agreement features resulted in non-significant overall improvements in three of the four models, but actually hurt the performance of the matrix level pronouns in the entity-mention models.

Table 5.9: **Syntactic and ACE type features: Bikel entity-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	76	38	162	22	3	27	328
he-they	10	6	47	4	2	30	99
she	4	2	29	1	0	11	47
she-they	0	1	3	0	0	2	6
it	4	2	17	0	0	0	23
Total	94	49	258	27	5	70	
	77.7%	90.7%	78.7%	22.3%	9.3%	21.3%	
			401			102	503
			79.7%			20.3%	

Table 5.10: **Syntactic and ACE type features: Bikel mention-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	78	35	158	20	6	31	328
he-they	6	6	42	8	2	35	99
she	4	1	32	1	1	8	47
she-they	0	1	2	0	0	3	6
it	3	2	17	1	0	0	23
Total	91	45	251	30	9	77	
	75.2%	83.3%	76.5%	24.8%	16.7%	23.5%	
			387			116	503
			76.9%			23.1%	

Table 5.11: **Syntactic and ACE type features: Xue entity-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	74	35	166	20	6	27	328
he-they	8	8	46	2	4	31	99
she	4	1	31	1	0	10	47
she-they	0	1	3	0	0	2	6
it	2	5	16	0	0	0	23
Total	88	50	262	23	10	70	
	79.3%	83.3%	78.9%	20.7%	16.7%	21.1%	
			400			103	503
			79.5%			20.5%	

Table 5.12: **Syntactic and ACE type features: Xue mention-mention**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	75	36	163	19	5	30	328
he-they	4	8	42	6	4	35	99
she	3	1	33	2	0	8	47
she-they	0	1	3	0	0	2	6
it	1	5	16	1	0	0	23
Total	83	51	257	28	9	75	
	74.8%	85.0%	77.4%	25.2%	15.0%	22.6%	
			391			112	503
			77.7%			22.3%	

Table 5.13: **Syntactic, agreement, and ACE type: Bikel entity-mention**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	74	37	160	24	4	29	328
he-they	8	7	49	6	1	28	99
she	4	2	31	1	0	9	47
she-they	0	1	3	0	0	2	6
it	4	2	16	0	0	1	23
Total	90	49	259	31	5	69	
	74.4%	90.7%	79.0%	25.6%	9.3%	21.0%	
			398			105	503
			79.1%			20.9%	

Table 5.14: **Syntactic, agreement, and ACE type: Bikel mention-mention**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	78	33	154	20	8	35	328
he-they	9	7	44	5	1	33	99
she	4	2	33	1	0	7	47
she-they	0	1	2	0	0	3	6
it	3	2	17	1	0	0	23
Total	94	45	250	27	9	78	
	77.7%	83.3%	76.2%	22.3%	16.7%	23.8%	
			389			114	503
			77.3%			22.7%	

Table 5.15: **Syntactic, agreement, and ACE type: Xue entity-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	74	35	167	20	6	26	328
he-they	7	8	45	3	4	32	99
she	4	1	33	1	0	8	47
she-they	0	1	4	0	0	1	6
it	2	5	16	0	0	0	23
Total	87	50	265	24	10	67	
	78.4%	83.3%	79.8%	21.6%	16.7%	20.2%	
			402			101	503
			79.9%			20.1%	

Table 5.16: **Syntactic, agreement, and ACE type: Xue mention-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	74	34	162	20	7	31	328
he-they	6	9	42	4	3	35	99
she	4	1	34	1	0	7	47
she-they	0	1	4	0	0	1	6
it	1	5	16	1	0	0	23
Total	85	50	258	26	10	74	
	76.6%	83.3%	77.7%	23.4%	16.7%	22.3%	
			393			110	503
			78.1%			21.9%	

5.6 Syntactic Plus Semantic Features, With and Without Agreement Features

Instead of using the ACE type constraint for semantics, semantic features analogous to the semantic feature checking in the Hobbs implementation (see pages 78*ff*) were implemented to see how they would compare to the hand-coded ACE types. The same nouns database was used, and a verbs database was constructed from the machine parses for the documents in the CACE training data in the same way that the one for the CTB was constructed.

Very simple "NounMatch" features were generated when the pronoun was the subject of a verb. If the proposed antecedent's noun category matched one or all of the verb's subject categories, then a feature "NounMatch:part" or "NounMatch:all" was generated, otherwise the feature was "NounMatch:none".

Given the non-robust method of creating the databases, and the rather on-the-fly nature of the features, it was encouraging to see that adding the features resulted in a slightly positive (although not significant) improvement over the syntax-only models (from 69.0% to 70.4% in Tables 5.1 and 5.17, and from 64.6% to 65.0% in Tables 5.2 and 5.18) for the models trained on the Bikel parses.

On the other hand, the two types of models trained on the Xue parses both decreased when the NounMatch features were added (although not significantly).

Adding agreement features to the semantic features improved the scores for the entity models but not both mention models, and not significantly.

Table 5.17: **Syntactic and semantic features: Bikel entity-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	74	38	148	24	3	41	328
he-they	8	4	31	6	4	46	99
she	4	1	26	1	1	14	47
she-they	0	1	4	0	0	1	6
it	1	2	12	3	0	5	23
Total	87	46	221	34	8	107	
	71.9%	85.2%	67.4%	28.1%	14.8%	32.6%	
			354			149	503
			70.4%			29.6%	

Table 5.18: **Syntactic and semantic features: Bikel mention-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	63	32	140	35	9	49	328
he-they	5	4	26	9	4	51	99
she	2	1	31	3	1	9	47
she-they	0	1	4	0	0	1	6
it	3	2	13	1	0	4	23
Total	73	40	214	48	14	114	
	60.3%	74.1%	65.2%	39.7%	25.9%	34.8%	
			327			176	503
			65.0%			35.0%	

Table 5.19: **Syntactic and semantic features: Xue entity-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	72	33	142	22	8	51	328
he-they	6	7	32	4	5	45	99
she	4	1	29	1	0	12	47
she-they	0	1	3	0	0	2	6
it	2	3	10	0	2	6	23
Total	84	45	216	27	15	116	
	75.7%	75.0%	65.1%	24.3%	25.0%	34.9%	
			345			158	503
			68.6%			31.4%	

Table 5.20: **Syntactic and semantic features: Xue mention-mention**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	58	33	138	36	8	55	328
he-they	1	6	27	9	6	50	99
she	2	1	32	3	0	9	47
she-they	0	1	4	0	0	1	6
it	1	5	12	1	0	4	23
Total	62	46	213	49	14	119	
	55.9%	76.7%	64.2%	44.1%	23.3%	35.8%	
			321			182	503
			63.8%			36.2%	

Table 5.21: **Syntax, Agreement, and Semantics: Bikel entity-mention**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	72	35	147	26	6	42	328
he-they	9	6	36	5	2	41	99
she	4	2	31	1	0	9	47
she-they	0	1	3	0	0	2	6
it	2	2	11	2	0	6	23
Total	87	46	228	34	8	100	
	71.9%	85.2%	69.5%	28.1%	14.8%	30.5%	
			361			142	503
			71.8%			28.2%	

Table 5.22: **Syntax, Agreement, and Semantics: Bikel mention-mention**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	60	28	136	38	13	53	328
he-they	6	5	34	8	3	43	99
she	3	2	30	2	0	10	47
she-they	0	1	3	0	0	2	6
it	2	2	14	2	0	3	23
Total	71	38	217	50	16	111	
	58.7%	70.4%	66.2%	41.3%	29.6%	33.8%	
			326			177	503
			64.8%			35.2%	

Table 5.23: **Syntactic, agreement, and Semantics: Xue entity-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	70	33	146	24	8	47	328
he-they	6	8	36	4	4	41	99
she	4	1	32	1	0	9	47
she-they	0	1	4	0	0	1	6
it	2	3	10	0	2	6	23
Total	82	46	228	29	14	104	
	73.9%	76.7%	68.7%	26.1%	23.3%	31.3%	
			356			147	503
			70.8%			29.2%	

Table 5.24: **Syntactic, agreement, and Semantics: Xue mention-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	59	26	139	35	15	54	328
he-they	2	8	36	8	4	41	99
she	3	1	28	2	0	13	47
she-they	0	1	4	0	0	1	6
it	1	4	14	1	1	2	23
Total	65	40	221	46	20	111	
	58.6%	66.7%	66.6%	41.4%	33.3%	33.4%	
			326			177	503
			64.8%			35.2%	

5.7 "Pragmatic" Features, Alone and in Combinations

Two additional feature types were used. The first type is really another semantic feature, while the second is a document-level feature that is labeled, for lack of a better term, a "pragmatic" feature.

The semantic feature uses the categories from the Rocling dictionary[5]. If the potential antecedent is a nominal mention (as opposed to another pronoun or a name) then a feature is created from the Rocling category for the the head noun of the nominal mention.

The "pragmatic" feature is simply the prefix of the document name of the text being processed. This was chosen to reflect the differences in usage between the newswire sources and the broadcast news sources.

When these features were used in combination with just the syntactic features (Tables 5.25-5.28), the performance was significantly better than syntactic features alone, better than syntactic plus agreement features only, and better than syntactic plus the NounMatch semantic features.

The summary tables Tables 5.30, 5.31, 5.32, and 5.33 give the bottom-line scores for each of the four models built from each feature combination. Rows labeled **c/C**, **d/D**, **g/G**, **h/H**, **k/K**, and **m/M** show the scores when these two features bundled under the "pragmatic" label are used.

5.8 Ablation Tests of Syntactic Features

As noted at the start, the maximum entropy model with just syntactic/surface features did not show the large difference in accuracy between matrix-level pronouns

[5] **http://www.aclclp.org.tw**. The categories were combined into coarser-grained groups than are found in the original dictionary.

Table 5.25: **Syntactic and "Pragmatic" features: Bikel entity-mention**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	74	38	149	24	3	40	328
he-they	9	5	31	5	3	46	99
she	4	2	32	1	0	8	47
she-they	0	1	4	0	0	1	6
it	2	2	12	2	0	5	23
Total	89	48	228	32	6	100	
	73.6%	88.9%	69.5%	26.4%	11.1%	30.5%	
			365			138	503
			72.6%			27.4%	

Table 5.26: **Syntactic and "Pragmatic" features: Bikel mention-mention**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	65	33	148	33	8	41	328
he-they	6	3	27	8	5	50	99
she	2	2	32	3	0	8	47
she-they	0	1	3	0	0	2	6
it	3	2	13	1	0	4	23
Total	76	41	223	45	13	105	
	62.8%	75.9%	68.0%	37.2%	24.1%	32.0%	
			340			163	503
			67.6%			32.4%	

Table 5.27: **Syntactic and "Pragmatic" features: Xue entity-mention**

	Correct			Wrong			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	74	33	154	20	8	39	328
he-they	8	7	32	2	5	45	99
she	4	1	30	1	0	11	47
she-they	0	1	3	0	0	2	6
it	1	4	10	1	1	6	23
Total	87	46	229	24	14	103	
	78.4%	76.7%	69.0%	21.6%	23.3%	31.0%	
			362			141	503
			72.0%			28.0%	

Table 5.28: **Syntactic and "Pragmatic" features: Xue mention-mention**

	Correct			**Wrong**			
	M	**M2**	**S**	**M**	**M2**	**S**	**Total**
he	61	32	145	33	9	48	328
he-they	3	5	28	7	7	49	99
she	2	1	33	3	0	8	47
she-they	0	1	3	0	0	2	6
it	1	5	12	1	0	4	23
Total	67	44	221	44	16	111	
	60.4%	73.3%	66.6%	39.6%	26.7%	33.4%	
			332			171	503
			66.0%			34.0%	

and subordinate-level pronouns that the rule-based method did.

To see if it could be determined which of the features were contributing to the performance of the different levels, a series of tests was performed using just one of the four model choices. Since the entity-mention model trained on the Bikel parses had the nominally highest overall score (trained with the ACE type constraint plus agreement, in Table 5.13), that combination of model type and parses was arbitrarily chosen for the tests.

Table 5.29 shows the scores for the entity-mention model trained on the Bikel parses using all of the features discussed in Section 5.3.1, but withholding the feature listed in the leftmost column.

Almost all of the features caused the score for the subordinate pronouns to go down when they were omitted, with "ClauseDepth" showing the largest drop in percentage when it was removed. It also contributes at the matrix level to some degree as well.

Only "PositionPair" and "HeadWords" resulted in improved scores for subordinate pronouns when they were left out.

While the feature "HeadWords" for the pronoun might seem to be redundant with the pronoun itself, in machine parses it captures those cases in which the parse is bad and the pronoun is combined with one or more other characters under a

different part-of-speech label (captured in the feature "HeadPOS"). This is probably more likely to happen in nested clauses than at the matrix level. The much greater variability of the actual words probably makes "HeadWords" too sparse to be useful, while "HeadPOS" generalizes to a category label that will occur more frequently, and not just in the training set.

"PositionPair" seems to contribute most at the matrix level relative to the other features, but improves the **S**-level pronoun score by its absence. The features "ConsecutiveSentenceStart", "SameSentence", "ExactMatchHead", and "MinMentionDist" behave in the opposite way, helping the subordinate level but hurting the matrix pronoun score.

It would be worthwhile to try experiments that condition these six features on the syntactic level or else combine the feature with the level to better exploit the features' contributions. An experiment on a different baseline with the "MinMentionDist" feature, for example, showed that only generating the feature when the pronoun is an **S**-level pronoun increased the scores for all levels over the base.

5.9 Summary

Tables 5.30, 5.31, 5.32, and 5.33 summarize the results from the two model types trained on the two sets of parses.

Of all the pronoun coreference resolvers presented in this thesis, the best performing on all levels combined are the entity-mention and mention-mention maximum entropy models that used the basic syntactic features and constrained the antecedent so that it could not be a different ACE type from the pronoun. Adding agreement features in the entity-mention models improved the **S**-level score, but reduced the **M**-level, while in the mention-mention models the addition of agreement helped at both levels.

Of particular interest is that the same maximum entropy models were able to

Table 5.29: **Ablation Tests of Syntactic/Surface Features**

Feature withheld	**Percent Correct and Counts**			
	M (121)	**M2** (54)	**S** (328)	**All** (503)
Base with all features	**69.4%** 84	**83.3%** 45	**66.5%** 218	**69.0%** 347
- Mentions.count	**68.6%** 83	83.3% 45	66.5% 218	**68.8%** 346
- ExactMatchExtent	**67.8%** 82	**77.8%** 42	**64.9%** 213	**67.0%** 337
- ExactMatchHead	*70.2%* 85	83.3% 45	**65.5%** 215	**68.6%** 345
-Consecutive SentenceStart	*70.2%* 85	**79.6%** 43	**65.2%** 214	**68.0%** 342
- SameSentence	*72.7%* 88	**79.6%** 43	**65.5%** 215	**68.8%** 346
- MinMentionDist	*70.2%* 85	**79.6%** 43	**64.9%** 213	**67.8%** 341
- MinCharDist	69.4% 84	**79.6%** 43	**65.2%** 214	**67.8%** 341
- HeadPOS	**68.6%** 83	**79.6%** 43	**64.9%** 213	**67.4%** 339
- HeadWords	69.4% 84	**81.5%** 44	*66.8%* 219	69.0% 347
- PositionPair	**66.9%** 81	**77.8%** 42	*68.9%* 226	*69.4%* 349
- Position	**68.6%** 83	83.3% 45	66.5% 218	**68.8%** 346
- CCommanded	**68.6%** 83	**77.8%** 42	**65.9%** 216	**67.8%** 341
- ClauseDepth	**68.6%** 83	**79.6%** 43	**64.3%** 211	**67.0%** 337
- TreeDepth	69.4% 84	**79.6%** 43	**64.9%** 213	**67.6%** 340
- VCPattern	69.4% 84	**79.6%** 43	66.5% 218	**68.6%** 345

Table 5.30: **Maxent: Experiments – Bikel Parses – Entity-mention**

Whodunit		Percent Correct M	M2	S	All
Syntax only	a	**69.4%**	83.3%	**66.5%**	**69.0%**
+ Agr	b	**67.8%**	83.3%	**69.2%**	**70.4%**
+ Prag	c	***73.6%***	88.9%	**69.5%**	**72.6%**
+ Agr + Prag	d	**72.7%**	85.2%	**69.2%**	**71.8%**
+ Sem	e	**71.9%**	85.2%	**67.4%**	**70.4%**
+ Sem + Agr	f	**71.9%**	85.2%	**69.5%**	**71.8%**
+ Sem + Prag	g	**76.0%**	85.2%	**69.5%**	**72.8%**
+ Sem + Agr + Prag	h	**74.4%**	85.2%	**70.1%**	**72.8%**
+ ACE type	i	**77.7%**	90.7%	**78.7%**	**79.7%**
+ ACE + Agr	j	**74.4%**	90.7%	**79.0%**	**79.1%**
+ ACE + Prag	k	**78.5%**	90.7%	**78.4%**	**79.7%**
+ ACE + Agr + Prag	m	**77.7%**	88.9%	**78.0%**	**79.1%**

resolve subordinate-level pronouns with an accuracy ranging from 76.2% (Table 5.31 row J) to to 79.8% (Table 5.32 row j), approximately double the best of the rule-based Hobbs algorithm for these difficult cases.

Observing differential effects of adding agreement features, and noting the opposite effects of the same feature on **S**-level pronouns *vs.* on **M**-level pronouns, however, leads us to suggest that the next logical step would be to train separate models for the two, or possibly three, levels in order to achieve the best scores for each kind of pronoun without compromise.

Table 5.31: **Maxent: Experiments – Bikel Parses – Mention-mention**

Whodunit		Percent Correct M	M2	S	All
Syntax only	A	**60.3%**	70.4%	**65.2%**	**64.6%**
+ Agr	B	**58.7%**	70.4%	**64.6%**	**63.8%**
+ Prag	C	**62.8%**	75.9%	**68.0%**	**67.6%**
+ Agr + Prag	D	**62.0%**	66.7%	**67.1%**	**65.8%**
+ Sem	E	**60.3%**	74.1%	**65.2%**	**65.0%**
+ Sem + Agr	F	**58.7%**	70.4%	**66.2%**	**64.8%**
+ Sem + Prag	G	**59.5%**	75.9%	**67.7%**	**66.6%**
+ Sem + Agr + Prag	H	**62.8%**	74.1%	**67.1%**	**66.8%**
+ ACE type	I	**75.2%**	83.3%	**76.5%**	**76.9%**
+ ACE + Agr	J	**77.7%**	83.3%	**76.2%**	**77.3%**
+ ACE + Prag	K	**76.9%**	85.2%	**76.2%**	**77.3%**
+ ACE + Agr + Prag	M	**76.9%**	85.2%	**75.0%**	**76.5%**

Table 5.32: **Maxent: Experiments – Xue Parses – Entity-mention**

Whodunit		Percent Correct M	M2	S	All
Syntax only	a	**73.0%**	75.0%	**66.3%**	**68.8%**
+ Agr	b	**74.8%**	73.3%	**69.0%**	**70.8%**
+ Prag	c	**78.4%**	76.7%	**69.0%**	**72.0%**
+ Agr + Prag	d	**76.6%**	78.3%	**70.5%**	**72.8%**
+ Sem	e	**75.7%**	75.0%	**65.1%**	**68.6%**
+ Sem + Agr	f	**73.9%**	76.7%	**68.7%**	**70.8%**
+ Sem + Prag	g	**79.3%**	73.3%	**69.6%**	**72.2%**
+ Sem + Agr + Prag	h	**77.5%**	76.7%	**71.1%**	**73.2%**
+ ACE type	i	**79.3%**	83.3%	**78.9%**	**79.5%**
+ ACE + Agr	j	**78.4%**	83.3%	**79.8%**	**79.9%**
+ ACE + Prag	k	**82.0%**	81.7%	**76.5%**	**78.3%**
+ ACE + Agr + Prag	m	**79.3%**	83.3%	**78.9%**	**79.5%**

Table 5.33: **Maxent: Experiments – Xue Parses – Mention-mention**

Whodunit		Percent Correct M	M2	S	All
Syntax only	A	**55.9%**	78.3%	**64.5%**	**64.2%**
+ Agr	B	**58.6%**	65.0%	**66.0%**	**64.2%**
+ Prag	C	**60.4%**	73.3%	**66.6%**	**66.0%**
+ Agr + Prag	D	**62.2%**	65.0%	**67.5%**	**66.0%**
+ Sem	E	**55.9%**	76.7%	**64.2%**	**63.8%**
+ Sem + Agr	F	**58.6%**	66.7%	**66.6%**	**64.8%**
+ Sem + Prag	G	**60.4%**	71.7%	**68.1%**	**66.8%**
+ Sem + Agr + Prag	H	**63.1%**	68.3%	**68.1%**	**67.0%**
+ ACE type	I	**74.8%**	85.0%	**77.4%**	**77.7%**
+ ACE + Agr	J	**76.6%**	83.3%	**77.7%**	**78.1%**
+ ACE + Prag	K	**76.6%**	85.0%	**78.3%**	**78.7%**
+ ACE + Agr + Prag	M	**76.6%**	83.3%	**77.4%**	**77.9%**

Chapter 6

Conclusions and Future Work

6.1 Introduction

In Chapter 1 we stated the following goals.

First, to demonstrate that a simple, rule-based algorithm, represented by the Hobbs algorithm, could perform reasonably well in resolving overt third-person pronouns in Chinese.

Second to show that the same algorithm could resolve matrix-level zero pronouns as well as overt pronouns.

Third, to test the hypothesis that richer semantic features obtained automatically from existing linguistic resources could be used as semantic constraints on the antecedents of pronominal anaphors to improve the performance of the Hobbs algorithm.

Fourth, to demonstrate that a maximum entropy, machine-learning model can be trained that will achieve a good performance in pronominal anaphora resolution in Chinese, and that using richer linguistic resources for features improves that performance.

After analyzing data from the first goal we discovered that the results corroborated work by Miltsakaki [33]. As a consequence, in the evaluation of the results

from each step, we analyzed the success of each method or model with respect to the syntactic level of the anaphor.

6.2 Summary of Results and Implications

The first contribution of this study was to provide a corpus of Chinese gold-standard parses annnotated for pronominal coreference. This was done by developing the guidelines that were discussed in Chapter 3, and with *much* help from native Chinese-speaking annotators who used the guidelines to annotate the parses of the Penn Chinese Treebank.

Using early versions of the annotated data we were able to demonstrate the following.

In Chapter 4 we showed that the syntax-only version of the Hobbs algorithm did not perform very well for all third-person pronouns, achieving only a combined accuracy of 57.1% in the CTB implementation and just 41.2% to 40.5% correct for all pronouns in the two sets of parses in CACE (Bikel and Xue, respectively). When the matrix-level pronouns alone were scored, however, the scores were much better, namely 70.1% for the CTB Hobbs, and 53.7% and 54.1% for the algorithm executed on the Bikel and Xue parses.

As for the zero pronouns, the syntax-only version of the algorithm did not perform nearly as well, with only 43.0% correct overall, and performance on matrix-level ***pro*** was even worse, at 40.2%. When semantic constraints were added, they were too limited and too noisy to improve the performance by more than one pronoun.

It is very likely that, as Kameyama [27] explained in detail for Japanese and stated about Chinese, that the zero pronouns are playing different discourse roles from the overt pronouns. In addition, it is assumed that these roles are conditioned primarily by semantic, pragmatic, and discourse criteria, not structural ones. This would explain the lack of success of an algorithm based on parsed structure.

Adding agreement and semantics to the syntax-only version of the Hobbs algorithm did indeed improve its performance on overt third-person pronouns in the CTB parses overall, but had different effects for pronouns at different syntactic levels. Both agreement and semantics helped the performance on **S** level pronouns more than on matrix-level pronouns, while semantics alone actually hurt the accuracy of resolving matrix-level pronouns in CTB. The "semantic" information in the ACE type was used for the CACE application, and had positive effects for all three syntactic levels. This was attributed to the fact that the ACE types are precise, hand-coded features, in contrast to the noisy automatically generated semantic databases.

The performance of the algorithm on the machine parses was not as good overall as the performance on the gold-standard parses. A comparison of the algorithm run on the same small set of texts showed a significant difference (P = 0.0055) between the scores for the gold standard CTB parses *vs.* the parses produced by the Bikel parser. The total accuracy in the CTB parses was 65.4% while the total accuracy in the CACE machine parses was 56.7% (Tables 4.17 and **??**).

The higher scores for this small subset compared to the two large datasets were probably due to the higher proportion of pronouns in the subset that were at the matrix level (53.9%), compared to the full CTB (48.4%) or full CACE datasets (24.′ the Bikel parses).

The two maximum-entropy model types that were trained on the two sets of parses performed much better on the CACE texts than the Hobbs algorithm did, for all combinations of constraints and features. While the Hobbs algorithm never had a combined success rate over 60.1%, the highest scores for a maximum entropy model were 79.9% overall, and 82.0% for matrix pronouns[1].

Moreover, while the highest score the rule-based Hobbs algorithm achieved for resolving pronouns in subordinate constructions was 50.8% (for the CTB syntax + agreement + semantics model, Table 4.27), the maximum entropy models were able

[1]Not in the same model, alas.

to reach quite respectable rates of accuracy, with the best obtaining 79.8% correct for the **S** level pronouns (Table 5.15: entity-mention model trained on the Xue parses with both the ACE type constraint and agreement).

As was the case for the Hobbs experiments, different sources of information affected the performance of the maxent models to different degrees for pronouns at different syntactic levels. Adding semantic information, as hypothesized, helped improve the scores of the pronouns in subordinate contexts to a greater degree than it improved the accuracy on matrix pronouns (and sometimes the noisy semantics even interfered at the matrix level).

When the ablation tests were peformed, however, some of the features showed results that unexpectedly conflicted with the reasons that they had been selected in the first place. For example, "ConsecutiveSentenceStart" was a feature that was expected to essentially emulate the behavior of the Hobbs algorithm's step 4: when a pronoun is the matrix pronoun of a sentence it is frequently at the start of a sentence, and it is likely that its antecedent is at the beginning of the previous sentence. Yet when this feature was removed, the score for the matrix pronouns increased instead of getting worse. Similarly "PositionPair" might have captured some of the inter-sentential relation between a subordinate pronoun and an antecedent earlier in the sentence, but turned out to only be contributing to the matrix level pronoun resolution.

On the other hand, when the "CCommanded" feature (which relation is handled by steps 2 and 3 in the Hobbs algorithm), the expected decrease in performance was seen at both subordinate and matrix levels. "SameSentence" is another feature like "PositionPair" that was expected to help with subordinate level pronouns, and was found to be doing so, but at the cost of hurting the matrix level.

Other than these kinds of combined features, there was not a direct mapping between the surface features used in the "syntax-only" model and the structures the Hobbs algorithm was examining at each step. To emulate Hobbs would require more

combined features, and apparently with more precision than ones like "PositionPair".

The finding that using ACE type in lieu of semantic information obtained in other ways improved the performance of both the rule-based and the statistical approaches is a useful result. Even though the ACE types used here, like the CTB parses, were from hand-annotated, "perfect" information, and the outputs of automatic taggers would not be as precise, the nature of the information sourse is less subject to noise than sources such as noun-category databases.

A finding that has been discussed in passing, but deserves highlighting, is the effect the syntactic distribution of overt pronouns has on the results of *any* method of pronoun resolution. All pronouns are not created equal. A method such as the Hobbs algorithm, which is good at finding antecedents for matrix-level pronouns, will have a high overall score when the corpus happens to have a high percentage of such pronouns relative to pronouns at subordinate levels. Researchers have observed different reflections of this, for example, in counting numbers of intra- *vs.* inter-sentential anaphor-antecedent links. But the **S** *vs.* **M** distinction might be a more useful way of looking at the data.

6.3 Future Work

When the experiments in Chapter 4 were started, only the first 325 files, 100K words, of the CTB were annotated for coreference. Now that the first 250K words have been annotated, a maximum entropy model to find the antecedents for zero pronouns is worth pursuing. There are approximately 2900 ***pro***s in total in the 9500 sentences of the 840 texts of CTB-3. Of these, just about half are labeled with categories and half are coindexed.

This work identified zero pronouns using an explicit string in the gold-standard parses of the CTB. The motivation was to test the performance of the Hobbs algorithm on the null pronouns independently of the noise of the task of finding them.

Yeh ([54]) only identified 56.4% of the zero pronouns in Chinese, while Ferrández *et al.* ([13]) were successful in correctly identifying 88 percent of the zeroes in their Spanish texts, which have more morphology for number and gender.

While the usual approach is to use heuristics based on expected syntactic and predicate structures to locate dropped subjects, the annotated CTB now provides a corpus for training a statistical model on the task.

Related to automatically detecting and resolving zero pronouns, and undoubtedly something that would be helpful in doing those tasks successfully, is an analysis of the zero pronouns with respect to the roles they are playing in the texts and how they differ from the overt pronouns.

The ACE type was used as a proxy for semantic information in the CACE dataset for both Hobbs and maximum entropy models. Having demonstrated the power of the hand-annotated information, it would be of interest to see how well the algorithm or models perform using automatically typed mentions, to get a more realistic comparision with the semantic databases.

The findings of the experiments using maximum entropy models that were presented in Chapter 5 suggest several future directions/experiments.

While we tried some substitution tests (not reported here), swapping in modified features for some of the core surface features, there is another kind of experiment we did not do that would be worth trying. It is the complement of an ablation experiment, namely to start with just one feature and see how well a model trained with it could do[2]

It is clear from the ablation studies in Section 5.8 that the features that were used in the basic "syntactic" models were sometimes divided in terms of their contribution to the overall score of the model. That is, a feature such as "SameSentence" helped at the subordinate level but detracted from the score at the matrix level. Other features such as the agreement features showed the same kind of inverse behavior as

[2]Thanks to Candy Sidner for this idea.

well, depending on the model.

There are at least two ways to use this information to "tune" models for pronoun resolution. One approach is to adjust the features themselves by making them conditional on syntactic level, or combining the feature with the syntactic level. Another approach is to just train separate models for the different syntactic levels.

We prefer the latter, because it allows a more focussed approach toward modeling the linguistic behavior of pronouns at the different syntactic levels. These results have corroborated Miltsakaki's observations (in [33]), and there are features that her work suggests that would be appropriate only for pronouns that are nested within subordinate clauses.

Since each experiment raises new questions, these ideas for ongoing research should suffice, **ad infinitum**.

Bibliography

[1] Chinatsu Aone and Scott William Bennett. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 122–129, Massachusetts Institute of Technology,
Cambridge, MA, 26-30 June 1995.

[2] Breck Baldwin. CogNIAC: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45, Madrid, Spain, 1997. W97-1306.

[3] Hsin-Hsi Chen. The transfer of anaphors in translation. *Literary and Linguistic Computing*, 7(4):231–238, 1992.

[4] Jinying Chen and Martha Palmer. Chinese verb sense discrimination using an EM clustering model with rich linguistic features. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Barcelona, July 21-26 2004.

[5] Noam Chomsky. On binding. *Linguistic Inquiry*, 11:1–46, 1980.

[6] H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22:1–39, 1986.

[7] Susan Converse. Developing guidelines for the annotation of anaphors in the Chinese treebank. In Benjamin K. T'sou, Olivia O. Y. Kwong, and Tom B. Y Lai, editors, *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, pages 3–9, Academia Sinica, Taipei, Taiwan, August 2002.

[8] Susan Converse. Resolving pronominal references in Chinese with the Hobbs algorithm. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, JeJu Island, Korea, October 2005.

[9] Ido Dagan and Alon Itai. A statistical filter for resolving pronoun references. In Y. A. Feldman and Alfred Bruckstein, editors, *Artificial Intelligence and Computer Vision: Proceedings of the Seventh Israeli Conference*, pages 125–135. Elsevier Science Publishers, 1991.

[10] Deborah A. Dahl. Focusing and reference resolution in PUNDIT. In *Proceedings AAAI-86. Fifth National Conference on Artificial Intelligence*, volume 2, pages 1083–1088, Philadelphia, PA, August 11-15 1986.

[11] Christiane Fellbaum, editor. *WordNet. An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[12] Antonio Ferrández, Manuel Palomar, and Lidia Moreno. An Empirical Approach to Spanish Anaphora Resolution. *Machine Translation*, 14:191–216, 1999.

[13] Antonio Ferrández and Jesús Peral. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting of the Association fro Computational Linguistics (ACL'00)*, pages 166–172, Hong Kong, October 2000.

[14] Niyu Ge, John Hale, and Eugene Charniak. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170, University of Montreal, August 1998.

[15] Peter C. Gordon, Barbara J. Grosz, and Laura A. Gilliom. Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, 17:311–347, 1993.

[16] Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

[17] Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307, 1993.

[18] Lynette Hirschman and Nancy Chinchor. MUC-7 coreference task definition (version 3.0). In *MUC-7 Proceedings*. Science Applications International Corporation, 13 July 1997. (http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html).

[19] Jerry Hobbs. Pronoun resolution. Technical Report 76-2, Department of Computer Science, City College, City University of New York, 1976.

[20] Jerry Hobbs. Resolving pronoun references. *Lingua*, 44:311–338, 1978.

[21] C.-T. James Huang. On the distribution and reference of empty pronouns. *Linguistic Inquiry*, 5(4):531–574, 1984.

[22] C.-T. James Huang. Remarks on empty categories in chinese. *Linguistic Inquiry*, 18(2):321–337, 1987.

[23] C.-T. James Huang. Pro-drop in Chinese: A generalized control theory. In O. Jaeggli and K. Safir, editors, *The Null Subject Parameter*, pages 185–214. Kluwer Academic Publishers, Boston, 1989.

[24] C.-T. James Huang. Remarks on the status of the null object. In Robert Freidin, editor, *Principles and Parameters in Comparative Grammar*, chapter 3, pages 56–76. The MIT Press, Cambridge, MA, 1991.

[25] Yan Huang. *The Syntax and Pragmatics of Anaphora. A study with special reference to Chinese.* Cambridge University Press, Cambridge, 1994.

[26] Yan Huang. *Anaphora. A Cross-linguistic Approach.* Oxford University Press, Oxford, 2000.

[27] Megumi Kameyama. *Zero Anaphora: The Case of Japanese.* Linguistics, Stanford University, 1985.

[28] Andrew Kehler. Current theories of centering for pronoun interpetation: A critical evaluation. *Computational Linguistics*, 23(1):467–475, 1997.

[29] Christopher Kennedy and Branimir Boguraev. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 113–118, Copenhagen, Denmark, 1996.

[30] Shalom Lappin and Herbert J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

[31] Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos. A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 135–142, 2004.

[32] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[33] Eleni Miltsakaki. Toward an aposynthesis of topic continuity and intrasentential anaphora. *Computational Linguistics*, 28(3):319–355, 2002.

[34] Ruslan Mitkov. *Anaphora Resolution.* Longman, London, 2002.

[35] Tatsunori Mori, Mamoru Matsuo, and Hiroshi Nakagawa. Constraints and defaults on zero pronouns in Japanese instruction manuals. In *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 7–13, Madrid, Spain, 1997. W97-1302.

[36] Thomas S. Morton. Coreference for nlp applications. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 173–180, Hong Kong, October 2000.

[37] Masaki Murata, Hitoshi Isahara, and Makoto Nagao. Pronoun resolution in Japanese sentences using surface expressions and examples. In *Proceedings of the Workshop on Coreference and Its Applications*, pages 39–46, College Park, Maryland, 22 June 1999. W99-0206.

[38] Hiromi Nakaiwa. Automatic extraction of rules for anaphora resolution of Japanese zero pronouns from aligned sentence pairs. In *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, Madrid, Spain, 1997. W97-1304.

[39] Hiromi Nakaiwa. Automatic identification of zero pronouns and their antecedents within aligned sentence pairs. In *Proceedings of the Fifth Workshop on Very Large Corpora*, pages 127–141, 1997. W97-0114.

[40] Manuel Palomar, Antonio Ferrández, Lidia Moreno, Patricio Martínez-Barco, Jesús Peral, Maximiliano Saiz-Noeda, and Rafael Muñoz. An algorithm for anaphora resolution in Spanish texts. *Computational Linguistics*, 27(4):545–567, 2001.

[41] Candace L. Sidner. Focusing in the comprehension of definite anaphora. In Michael Brady and Robert C. Berwick, editors, *Computational Models of Discourse*, chapter 5, pages 267–330. MIT Press, Cambridge, MA, 1983.

[42] Zhiyi Song. A comparative study of subject pro-drop in Old Chinese and Modern Chinese. Linguistics, University of Pennsylvania, 2005.

[43] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–554, 2001.

[44] Michael Strube. Never look back: An alternative to centering. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 1251–1257, 1998.

[45] Joel R. Tetreault. Analysis of syntax-based pronoun resolution methods. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 602–605, 1999.

[46] Joel R. Tetreault. A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, 27(4):507–520, 2001.

[47] Marilyn Walker, Masayo Iida, and Sharon Cote. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):1–37, 1994.

[48] Marilyn A. Walker. Evaluating discourse processing algorithms. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 251–261, 1989.

[49] Marilyn A. Walker, Masayo Iida, and Sharon Cote. Centering in Japanese discourse. In *Proceedings of COLING*, Helsinki, 1990.

[50] Bonnie Lynn Webber. Structure and ostension in the interpretation of discourse deixis. *Natural Language and Cognitive Processes*, 6(2):107–135, 1991.

[51] Nianwen Xue and Martha Palmer. Automatic semantic role labeling for Chinese verbs. In *Proceedings of the 19th International*

Joint Conference on Artificial Intelligence, Edinburgh, Scotland, 2005. (**http://www.cis.upenn.edu/~chinese/cpb**).

[52] Nianwen Xue and Fei Xia. The Bracketing Guidelines for the Penn Chinese Treebank Project. Technical Report 00-08, Institute for Research in Cognitive Science, University of Pennsylvania, 2000. (**http://www.cis.upenn.edu/~chinese/parseguide.3rd.ch.pdf**).

[53] Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 10(4):1–30, 2004. (**http://www.cis.upenn.edu/~chinese**).

[54] Ching-Long Yeh and Yi-Chun Chen. Zero anaphora resolution in Chinese with shallow parsing. *Journal of Chinese Language and Computing*, to appear, 2005.