INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

An Approach To Anaphoric Pronouns

by

Niyu Ge

B.S. , Computer Science, Illinois Institute of Technology, 1994

M.S., Computer Science, Brown University, 1998

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the department of Computer Science at Brown University

Providence, Rhode Island

May, 2000

UMI Number: 9987763

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

This dissertation by Niyu Ge is accepted in its present form

by

the Department of Computer Science as satisfying the dissertation requirement

for the degree of Doctor of Philosophy.

Date May 9 2000

Eugene Charniak, Director

Recommended to the Graduate Council

Date 8th May 2000

Mark Johnson, Reader

Date 05/08/00

Tom Dean, Reader

Approved by the Graduate Council

Date 5/17/00

Peder J. Estrup

Dean of the Graduate School and Research

Vita

Vitals Niyu Ge was born on January 4, 1973 in Shanghai, People's Republic of China. In 1991, she came to the United States of America. After graduating from Illinois Institute of Technology, she worked in Panasonic Factory Automation for a year. She then went to pursue her graduate degrees in computer science. She now lives in Providence, RI.

Education

Brown University, Providence, RI

Ph.D. in Computer Science	May 2000 (expected)
M.S. in Computer Science	May 1998

Thesis: *An Approach to Anaphoric Pronouns*

Advisor: Professor Eugene Charniak

Illinois Institute of Technology, Chicago, IL

B.S. in Computer Science

Graduated cum laude, Class of 1994

GPA: 4.0

Honor: Dean's List all semesters

Awards

IGERT Scholarship Award, Brown University 1999

Outstanding Academic Achievement Award, Illinois Institute of Technology 1994

Outstanding Service Award, Illinois Institute of Technology 1994

Abstract

In recent years, the phenomenon of anaphora — where two linguistic expressions refer to the same discourse entity — has become an active study in both formal and computational linguistics. The correct interpretation of anaphoric pronouns is an important problem for many Natural Language Processing (NLP) applications such as information retrieval, natural language interfaces, machine translation, topic identification, and many more.

It has been observed that anaphora resolution involves syntactic, semantic, and pragmatic factors. While formal linguists have come to focus on the factors for determining disjoint reference between two expressions, i.e. which noun phrase cannot be the antecedent of an anaphoric pronoun, computational linguists have concentrated on finding the referential expression of an anaphoric pronoun. This thesis falls into the second category. There are two main objectives of this thesis: the first is to present a computational approach to resolve anaphoric pronouns in English text, and the second is to show the relative importance of the factors involved in anaphora resolution and proposes a core set of factors which are essential.

The anaphora resolution system we built is based on a probabilistic model. This model combines different linguistic evidence in a statistical framework. In contrast to many previous approaches, our system is completely automatic, uses a very small training set which does not require a large amount of manual marking, and achieves a very competitive success rate of 92.2%. From analyses of the components of our system and comparing it with other approaches, we have statistical evidence to believe that not all factors are equally important. Some factors form a core set which is essential in any approach to anaphora while others depend on the domain, the language, etc.

Acknowledgements

To my grandfather

There are so many people I want to thank. Most of all, I want to thank my committee. Eugene Charniak is an amazing advisor, who has given me wonderful support, guidance, and encouragement over the years. I have been extremely fortunate to have Mark Johnson on my committee. His linguistic work as well as his statistic knowledge has instrumental influence on me. Thanks also to Tom Dean for the almost thankless task of serving on my committee.

This thesis would be impossible without the support and love of my husband and my parents. In a sense, this thesis is for them.

In my second year at Brown, John Hale initiated the formation of a Natural Language Processing group which has been holding a weekly meeting since. The meetings have been very instructive and helpful. I am very grateful to John Hale, Brian Roak, Don Blaheta, Keith Hall, and indeed the whole NLP group for helpful discussion of many of the ideas presented in this thesis. To Julie Sedivy whose courses on semantics and anaphora have been most inspiring, I owe a special debt of gratitude.

I would not have come to Brown to pursue a Ph.D. degree doing NLP had it not been for Martha Evens who first aroused my fascination with the subject in her natural language course.

I would also like to thank NSF for providing most of my funding with NSF DGE 9870676 and the IGERT scholarship program (NSF Learning and Intelligent Systems grant SBR-9720368) which supported me in my last year at Brown.

Contents

Chapter 1 Introduction **1**

1.1 A syntactic account 1

1.2 A pragmatic/discourse account 4

1.3 Summary 8

Chapter 2 Previous Work **9**

2.1 Hobbs' algorithm 9

2.2 BFP centering algorithm 12

2.3 RAP: Resolution of Anaphora Procedure 14

2.4 Mitkov's algorithm 17

2.5 Conclusion 18

Chapter 3 A Statistical Approach **24**

3.1 Two probabilistic models 24

3.1.1 The basic model *24*

3.1.2 The syntactic-prominence model *33*

3.1.3 A special case *39*

3.1.3.1 The equation 40

3.1.3.2 The patterns 41

3.2 Inside the equations 42

3.2.1 Gender/Number/Animacy information – $P(\rho \mid W_a)$ *42*

3.2.2 Syntactic prominence – $P(G_w \mid G_\rho)$ *44*

3.2.3 Discourse salience – $P(a \mid M_a, S_\rho)$ and $\frac{P(M_a \mid a, S_\rho, G_\rho)}{P(M_a)}$ *46*

3.2.4 Sentence recency — $P(S_w \mid a, f_\rho)$ *48*

3.2.5 Syntactic constraints – $P(d_a \mid a, f_\rho)$ *49*

3.2.6 Lexical semantics – $\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$ *51*

3.2.7 Pleonastic pattern statistic – P(pattern | pleonastic) *53*

3.3 Implementing the algorithm 54

3.3.1 Implementing Hobbs' algorithm *54*

3.3.2 Collecting statistics *58*

3.3.3 Gathering more information *64*

3.3.3.1 Unsupervised learning of gender information 64
3.3.3.2 Near "perfect" information 66
3.3.3.3 Simple transductive learning technique 68
3.3.3.4 Adjectives for pleonastic IT 68
3.3.3.5 Bllip99 statistics 69
3.3.4 Resolving pronouns *70*
3.4 The experiment 74
3.4.1 Corpus annotation *74*
3.4.2 Empirical results *76*
3.4.3 Comparing the two models *80*
3.4.4 Classifying errors *84*
3.5 Comparison with previous approaches 92

Chapter 4 Factors In Anaphora Resolution 96

4.1 Pronouns themselves 96
4.2 Grammatical salience 98
4.3 Discourse salience 100
4.4 c-command 101
4.5 Lexical semantics 102
4.6 World knowledge/Context-based inference 104
4.7 Summary 105

Chapter 5 107

Further Applications in the Statistical Framework 107

5.1 Pronouns in text generation 107
5.2 Centering revisited: a statistical attempt 112

Chapter 6 Conclusion 115

6.1 Summary of the models 115
6.2 Future research 116

Appendix A Hobbs' Algorithm 121

Appendix B 124

Distributions of antecedents' grammatical roles 124

Appendix C Likelihood-ratio Test **125**

Appendix D **127**

Adjectives for Recognition of Pleonastics **127**

Bibliography **128**

List of Figures

Figure 1.1 c-command explained 2
Figure 1.2 c-command: reflexive 3
Figure 1.3 c-command: non-reflexive 4
Figure 2.1 Adjunctive PP 10
Figure 2.2 PP Head noun attachment 11
Figure 3.1 $P(d_a \mid a, f_p)$ 51
Figure 3.2 Example of S type 52
Figure 3.3 Example of VP type 52
Figure 3.4 Minimal domain 56
Figure 3.5 Collapsing a parse tree 57
Figure 3.6 Raising a parse tree 58
Figure 3.7 UMSBJ: case 1 60
Figure 3.9 ESBJ configuration 61
Figure 3.8 UMSBJ: case 2 62
Figure 3.10 NPSBJ configuration 63
Figure 3.11 PP example 63
Figure 3.12 PPS example 64
Figure 3.13 *EXP* empty nodes 69
Figure 3.14 The resolution system 71
Figure 3.15 Accuracy vs. Head frequency 80
Figure 3.16 Accuracy vs. Head Noun frequncy 80
Figure 3.17 Accuracy vs. Head Noun frequency: large corpus 82
Figure 3.18 Accuracy vs. Head Noun frequency: small and big corpus 83
Figure 5.1 P(c | M_c) 109
Figure 5.2 P(c | comp) 109

List of Tables

Table 1.1 Transitions in centering 7

Table 1.2 Transition table for 1.4c and 1.4c' 8

Table 2.1 Transitions in BFP centering algorithm 13

Table 2.2 BFP example 14

Table 2.3 Ordering of $Cf(U_{i-1})$ affects $Cb(U_i)$ 19

Table 2.4 Restrictiveness of Rule 1 20

Table 2.5 RAP's salience weighting 21

Table 3.1 Factors in the Basic model 33

Table 3.2 Factors in the Syntactic-prominence model 39

Table 3.3 Gender/Number/Animacy – $P(p \mid W_a)$ 43

Table 3.4 $P(G_w \mid G_p)$ – four categories 45

Table 3.5 $P(G_w \mid G_p)$ – subjecthood inheritance 45

Table 3.6 Mention counts $P(a \mid M_a, S_p)$ – used in the Basic model 46

Table 3.7 Mention counts $\frac{P(M_a \mid a, S_p, G_p)}{P(M_a)}$ — used in the Syntactic-prominence model 47

Table 3.8 Sentence recency – $P(S_w \mid a, f_p)$ 48

Table 3.9 Hobbs' distance – $P(d_a \mid a)$ 49

Table 3.10 Hobbs' distance – $P(d_a \mid a, f_p)$ 50

Table 3.11 Lexical semantics — $\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$ 53

Table 3.12 Pleonastic pattern – $P(pattern \mid pleonastic)$ 54

Table 3.13 Hale's good statistics 65

Table 3.14 Hale's noisy statistics 65

Table 3.15 WHICH and WHO 67

Table 3.16 WHICH and WHO overlaps 67

Table 3.17 Top 15 most likely words 72

Table 3.18 Singular collective nouns 73

Table 3.19 HE/SHE/IT results on test data 77

Table 3.20 Pleonastic *IT*s in test data 77

Table 3.21 Performance on the test data 78

Table 3.22 Cross validation results on HE/SHE/IT 78

Table 3.23 Incremental results – Basic model 79
Table 3.24 Incremental results – Syntactic-prominence model 79
Table 3.25 Error classification 85
Table 3.26 Evaluation of the centering algorithm on the incorrect output 90
Table 3.27 Centering algorithm applied to error output:correct 90
Table 3.28 Centering algorithm applied to error output: incorrect 91
Table 3.29 Centering algorithm applied to the correct output 91
Table 3.30 Comparison of the algorithms 104
Table 5.1 $P(G_e \mid c)$ 109
Table 5.2 $P(c \mid M_e)$ 109
Table 5.3 P(c | comp) 109

Chapter 1 Introduction

This thesis is organized as follows. A brief introduction is given in chapter 1. In chapter 2, I will survey four representative anaphora resolution programs. A detailed presentation of our statistical model is in chapter 3. In chapter 4, I will identify a core set of factors for anaphora resolution. Chapter 5 discusses two further applications in the framework and I will conclude in chapter 6.

In this chapter I will introduce different accounts of the anaphora problem. Anaphora is known to interact with various syntactic, semantic, and pragmatic considerations. I will briefly discuss some of these factors that are most relevant in a computational approach.

1.1 A syntactic account

A syntactic approach to anaphora focuses on the study of the intrasentential relationship between pronouns and full noun phrases (NP) within the framework of generative grammar. Important linguists like Langacker (1969), Chomsky (1980,1981) and Reinhart (1981,1983), to name just a few, have sought to explain this phenomenon on the basis of syntactic structure. The notion of c(onstitute)**-command** (Reinhart 1981, 1983) has a central role within this framework. The definition of **c-command** is the following:

Definition: Node A *c(onstitute)-commands* node B iff the branching node immediately dominating A also dominates B. (Reinhart 1981)

In the following tree structure,

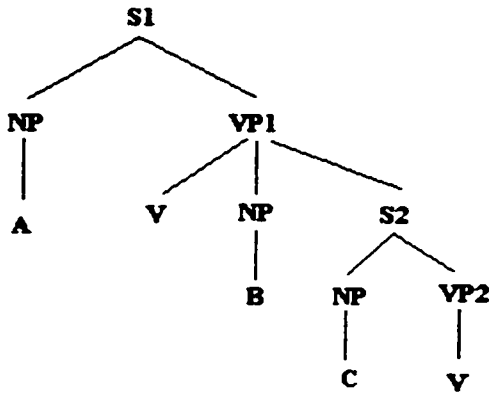


Figure 1.1 c-command explained

Node A c-commands the whole tree because its immediate branching node S_1 dominates all the other nodes in the tree. Node B only c-commands the elements under VP_1 and the c-command domain of node C is S_2.

Using English sentences as her data, Reinhart argues convincingly that the notion of c-command is the correct restriction on coreference between NPs (with a few exceptions). She formulates the constraints as follows (Reinhart 1981):

- A reflexive or reciprocal pronoun must be interpreted as coreferential with (and only with) a c-commanding NP within its minimal governing category.
- A given NP cannot be interpreted as coreferential with a distinct non-pronoun in its c-commanding domain.

The definition of a *governing category* is a bit involved. To put it simply, the governing category of a node X is the minimal node that contains X, X's governor, and an "accessible" subject. The exact definitions of governor and "accessible" subject are rather complicated and are not directly relevant to the purpose of this discussion. Readers interested in this topic are referred to (Haegeman 1991). For this discussion, *governor* can be thought of as the head of a constituent and *accessible subject* as the subject under an S node. An example will make this clearer. In the following parse tree:

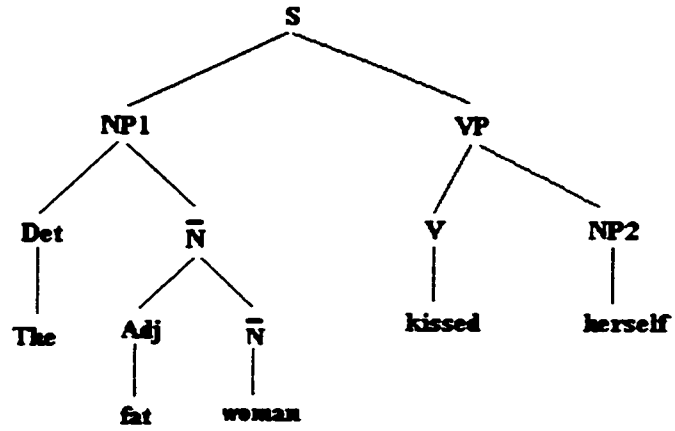


Figure 1.2 c-command: reflexive

The governor of NP_2 is the V(erb) node (the one dominating "*kissed*") since it is the head of the constituent VP. The node S contains NP_2 ("*herself*"), its governor V, and an accessible subject NP_1. Hence by definition, S is the governing category of NP_2. By the first binding principle, the NP_2 ("*herself*") must be interpreted as coreferential with an NP within its governing category S. This thus allows NP_1 ("the fat woman") to be the antecedent. If on the other hand, the "*herself*" is replaced by "*her*" as in Figure 1.3, the governing category of NP_2 ("*her*") is still S and by the second principle, the NP1("*the fat woman*") being in the governing category and c-commanding NP_2, cannot be coreferential with NP_2 ("*her*").

The GB framework has very elaborate and delicate theories on the restriction on coreference between pronouns and full noun phrases. The presentation here is much simplified. The important point is that the original intention of the anaphora question is to study the complementary distributions of reflexive/reciprocal pronouns (called **anaphors**) and regular pronouns (simply **pronoun**). In this thesis, I use the term *anaphora* to mean both and will use *reflexive/possessive/regular* to describe the pronoun if distinction is needed.

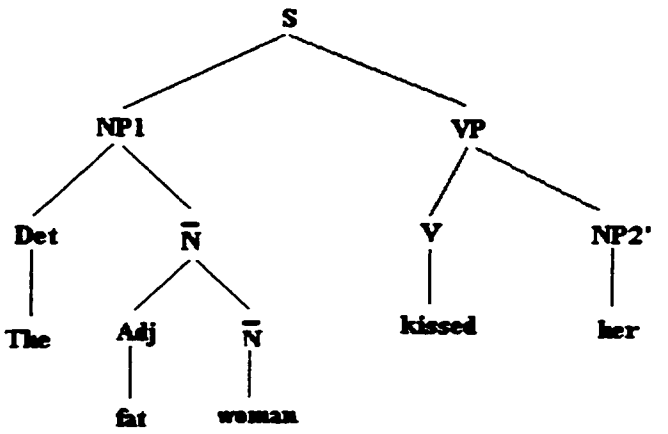


Figure 1.3 c-command: non-reflexive

1.2 A pragmatic/discourse account

In contrast to a pure syntactic account, a pragmatic approach looks at anaphora from the point of view of information relationships. This framework deals with language as a process of communication, and the organization of information is thus an important aspect of it. It holds that language processing must take into account the contextual structure of language, particularly with regard to reference, in order to constrain processes of inference and make them computationally tractable (Grosz 1977). **Centering theory** (Grosz et al. 1983, 1986, 1995) has been developed as an account of one aspect of discourse processing and local discourse structure that makes specific claims about both processing complexity and discourse anaphora. I will first discuss some of the motivations for this approach not only because they are important for an understanding of the theory but also they will, I believe, continue to motivate future research. The latter part of this section then presents the theory itself.

Centering theory is motivated by some peculiar linguistic phenomena that cannot be explained by purely content-based models of reference and coherence. It has long been observed

that pronouns and full noun phrases that corefer are not equivalent in terms of their information content. Lakoff (Lakoff 1968) set up an "anaphora hierarchy" which distinguishes between four types of noun phrases corresponding to the conditions under which these NPs can serve as anaphoric expressions: proper names, definite descriptions, epithets, and pronouns. NPs at different level of this hierarchy have different effect on coherence. Grice's (Grice 1975) conversational maxim on Quantity also predicts this difference:

Do not make your contribution more informative than required.

In other words, it is not necessary to repeat the full noun phrase if the reference of this NP is absolutely clear if a pronoun is used. This effect can be observed in a comparison of sentences (1.1c) and (1.1c'):

1.1a Susan gave Betsy a pet hamster.
b She reminded her that such hamster were quite shy.
c She asked Betsy whether she liked the gift.
c' Susan asked Betsy whether she liked the gift.

Psycholinguistic experiments showed that (1.1c') is more difficult to process than (1.1c) (Gordon 1993, Gordon & Chan 1995)

It is also been established that entities mentioned in a sentence (or an utterance) have different focus values and this in turn puts constraints on the use of pronouns. Compare (1.2c) and (1.2c'):

1.2a Jeff helped Dick wash the car.
b He washed the windows as Dick waxed the car.
c He soaped a pane.
c' He buffed the hood.

A purely semantic theory of discourse understanding cannot predict the difference in coherence because the "*He*" in (1.2c) can only cospecify with "*Jeff*" since the verb "*soap*" is related to "*wash*" whereas the "*He*" in (1.2c') can only refer to "*Dick*" since the verb "*buff*" is related to "*wax*". Centering, on the other hand, will predict the process difference. To put it simply, (1.2c) is a ***continuation*** of the discourse set up by (1.2a) and (1.2b) while (1.2c') causes a ***shift***. This will become clear when we discuss the theory which is the next topic.

In Grosz, Joshi, and Weinstein's (GJW 1995, henceforth **GJW**) centering theory, a discourse segment consists of a sequence of utterances $U_1,\ldots,U_n$. The set of **forward-looking centers Cf(U_i)** represents discourse entities realized in utterance U_i. This set is ranked according to discourse salience and the ranking is a partial order. The highest-ranked element on this list is called the **preferred center Cp(U_i)**. We can think of the elements in the Cf (U_i) list as candidates to be pronominalized in the next utterance. The **backward looking center Cb (U_i)** is a special member of Cf (U_{i-1}) representing the discourse entity that is most central in U_i. It links U_i to U_{i-1} and can be thought of as what U_i is "about". A key aspect of centering theory is the distinction between looking back to the previous discourse via Cb and predicting preferences for subsequent pronominalization via Cp. In addition to the structures of centers Cf, Cp, and Cb, there is a set of rules and constraints. (Gorden, Grosz, and Guillion 1993)

Constraints

For each utterance U_i in a discourse segment D consisting of utterances $U_1,\ldots U_n$:

1. There is precisely one backward-looking center Cb(U_i, D).
2. Every element of the forward-looking center's list Cf(U_i, D) must be realized[1] in U_i.
3. The center Cb(U_i, D) is the highest-ranked element of Cf(U_{i-1}, D) tat is realized in U_i.

Basically, these constraints say that every utterance has a central topic represented by Cb(U_i) and the ranking of the forward looking centers Cf determines from among the elements that are realized in the next utterance which of them will be the Cb for that utterance. Cf ranking is usually determined by the grammatical role in which an entity is realized[1]. There are also two rules proposed in GJW:

Rules

For each utterance U_i in a discourse segment D consisting of utterances $U_1,\ldots U_n$:

---

[1] This constraint depends on the definition of "realizes". The simplest definition is to take "realized" as "mentioned". For detailed discussions, see GJW86 and GJW95.

1. If some element of $Cf(U_{i-1}, D)$ is realized as a pronoun in U_i, then so is $Cb(U_i, D)$.
2. Transition states are ordered. The CONTINUE transition is preferred to the RETAIN transition, which is preferred to the SHIFT transition.

Rule 1 explains the oddness of (1.1c'). The example is reproduced here together with the centers for each sentence:

1.3a Susan gave Betsy a pet hamster.
Cf = {Susan, Betsy, hamster}; Cb = { }

b She reminded her that such hamsters were quite shy.
Cb = {Susan <realized by *She*>}; Cf = {Susan, Betsy <realized by *her*>, hamsters}

c She_1 asked Betsy whether she liked the gift.
Cb = {Susan <realized by She_1>}; Cf = {Susan, Betsy, hamster <realized by *the gift*>}

c' Susan asked Betsy whether she liked the gift.
Cb = {Susan}; Cf = {Susan, Betsy, hamster <realized by *the gift*>}

In (1.3c') the Cb{Susan} is not pronominalized where a non-Cb {Betsy} is. Rule 2 provides an ordering of transitions that can be used to measure coherence. Definitions of these transition types are given in Table 1.1.

	$Cb(U_i) = Cb(U_{i-1})$ or $Cb(U_{i-1}) = \{\}$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	CONTINUE	SHIFT
$Cb(U_i) \neq Cp(U_i)$	RETAIN	

Table 1.1 Transitions in centering

Rule 2 can be used to illustrate the difference between (1.2c) and (1.2c') reproduced here as (1.4c) and (1.4c'):

1.4a Jeff helped Dick washed the car.
Cb = { }; Cf = {Jeff, Dick, car}

b He washed the windows and Dick waxed the car.
Cb = {Jeff <realized by *He*>}; Cf = {Jeff, windows, Dick, car}

c He soaped a pane.

c' He buffed the hood.

---

[1] There are may proposals on Cf ranking. Some consider surface order of realization and some incorporate information status. But all rankings rely on grammatical roles. The ranking is also language dependent. See Kuno76, Kameyama 88, and Iida 92.

Utterance	Cb	Cp	Transition
1.4c	Jeff	Jeff	CONTINUE
1.4c'	Dick	Dick	SHIFT

Table 1.2 Transition table for 1.4c and 1.4c'

(1.4c') results in a SHIFT of discourse topic whereas (1.4c) results in a CONTINUE. This explains why (1.4c) is more coherent than (1.4c').

1.3 Summary

In this chapter I presented an overview of a purely syntactic approach to anaphora which will help the discussion of Hobbs' algorithm in the next chapter. We discussed the phenomena unexplained by purely content-based models of reference and coherence which motivate pragmatic/discourse-based models. Centering theory is such a model. In the next chapter I will also present an algorithm based on the centering theory.

Chapter 2 Previous Work

In this chapter, I will survey four existing algorithms for anaphora resolution. There are many other approaches in the literature. These four are chosen because of their representativeness of different accounts of anaphora. Hobbs' algorithm is a syntax based approach, BFP algorithm is inspired by the centering theory, Lappin and Leass' RAP system and the most recent work by Mitkov combine these various factors but with different emphasis.

2.1 Hobbs' algorithm

The Hobbs' algorithm (Hobbs 1976) is a syntactic approach and is based on traversing the parse trees of input sentences in a particular order looking for noun phrases of the correct gender and number. The algorithm incorporates the constraints on coreferentiality (i.e. disjoint reference) between a non-reflexive pronoun and a noun phrase.

The algorithm starts by looking for the antecedent within the current sentence in which the pronoun in question occurs (i.e. an intrasentential antecedent). It goes sequentially further and further up the tree to the left of the pronoun. In order to obey the syntactic constraints on coreference, the algorithm assumes that an NP node has an $\overline{N}$ node below it which denotes the noun phrase without its determiner and to which a prepositional phrase containing an argument of the head noun may be attached. This is distinguished from true adjunctive prepositional phrases which are attached to the NP node. This distinction is illustrated by the two examples in Figure 2.1 and 2.2. (Hobbs 1976)

This distinction is necessary in processing sentences (2.1) and (2.2):

(2.1) John saw a driver in his truck.
(2.2) John saw a driver of his truck.

In (2.1) the pronoun "*his*" may corefer with "*the driver*" whereas in (2.2) it may not. Recall that a non-reflexive pronoun cannot be interpreted as coreferential with a noun phrase in its governing domain (section §1.1). Having made this assumption, the algorithm implements the coreference constraint by skipping over NP nodes whose $\overline{N}$ node dominates the part of the parse tree in which the pronoun resides.

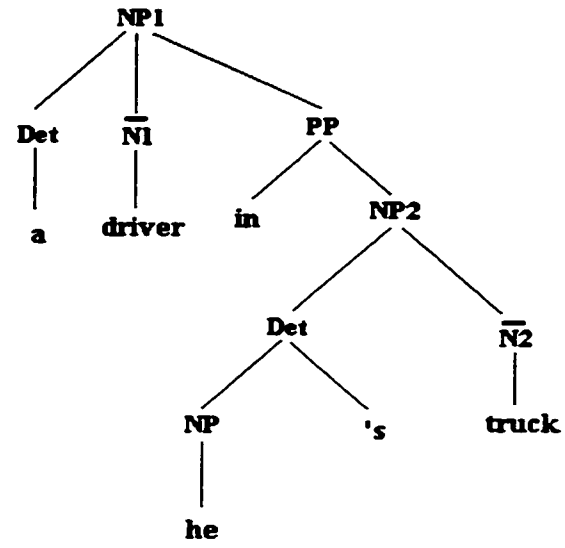


Figure 2.1 Adjunctive PP

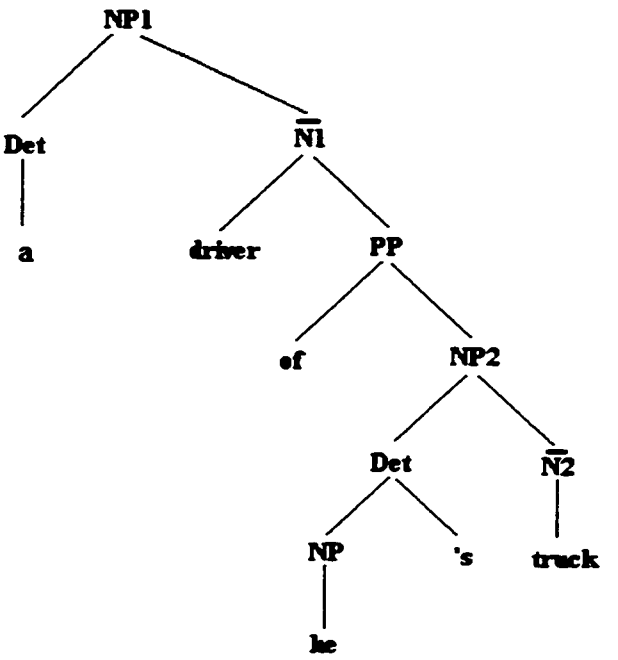


Figure 2.2 PP: head noun attachment

For the parse tree fragment in Figure 2.1, the algorithm goes up from the pronoun "*he*" to NP_2. Since there is no antecedent under NP_2 to the left of the pronoun "*he*", the search continues up the tree to NP_1. Note that the path leading to NP_1 ($NP_2 \rightarrow PP \rightarrow NP_1$) does not pass through the $\overline{N1}$ under NP_1. Hence NP_1 is a possible antecedent. In contrast, in Figure 2.2, we see that $\overline{N1}$ is on the path leading to NP_1 ($NP_2 \rightarrow PP \rightarrow \overline{N1} \rightarrow NP_1$). The constraint then is fired and NP_1 is ruled out. The algorithm continues in this fashion until it reaches the top S. If by this time no antecedent is found, the algorithm searches the previous sentence in a left-to-right BFS manner.

Though Hobbs' algorithm is a syntactic approach, the way it traverses the parse trees does in a way predict the salience of discourse entities. When searching for an antecedent in previous sentences (i.e. intersentential antecedent), the algorithm traverses the parse trees breadth-first, left-to-right. This amounts to giving entities realized in the subject position more salience and then objects and indirect objects. In general, the depth of tree embedding determines

discourse prominence. For example, even if an adjunct clause linearly precedes the main subject, since any NPs within it will be deeper in the parse tree, they would be predicted as less prominent than the subject. The algorithm is reproduced in full in the appendix.

For evaluation, the algorithm was hand simulated on three texts. The simulation was carried out with various assumed *knowledge*:

1. gender/number/person feature of a noun phrase is assumed known
2. selectional restriction is assumed known
3. the algorithm can somehow collect discourse entities into sets as coreferential with plural pronouns

An example of the last assumption is provided by sentence (2.3)

(2.1) John took Mary to a party. They had a lousy time.

The algorithm would "know" that "*John and Mary*", though not explicitly appearing in the sentence, is an antecedent for the pronoun "*They*". With items 1 and 3 above, Hobbs reports an average success rate of 88.3% (average of 85%, 88%, and 92%), and 91.7% (average of 92%, 90%, 93%) after using selectional restriction (item number 2). Marilyn Walker (Walker 1989) also manually evaluated the algorithm on three different sets of texts and she reported an average of 80%.[1]

2.2 BFP centering algorithm

The centering algorithm as proposed by Brennan, Friedman, and Pollard (Brennan, Friedman, and Pollard 1987, henceforth BFP), is based on the centering principles (see section §1.2 for detail). The algorithm utilizes Rule 1 to constrain the realization of centers and Rule 2 to order the movement of centers. Observing that there seem to be more and less coherent ways to *shift*, BFP proposes an extension to Rule 2, which handles some additional cases containing multiple ambiguous pronouns. The transition table used in BFP is presented in Table 2.1.

	$Cb(U_i) = Cb(U_{i-1})$ or $Cb(U_{i-1}) = \{\}$	$Cb(U_i) \neq Cb(U_{i-1})$
$Cb(U_i) = Cp(U_i)$	CONTINUE	SMOOTH-SHIFT
$Cb(U_i) \neq Cp(U_i)$	RETAIN	ROUGH-SHIFT

Table 2.1 Transitions in BFP centering algorithm

Intuitively the SMOOTH-SHIFT represents a more coherent way to *shift*. The preferred ranking is then CONTINUE ≻ RETAIN ≻ SMOOTH-SHIFT ≻ ROUGH-SHIFT where "≻" can be read as "is preferred to". The BFP centering algorithm consists of three phases:

1. CONSTRUCT all possible <Cb, Cf> pairs for the sentence.
2. FILTER the proposed pairs generated in step 1, e.g. by contra-indices, centering constraints and rules, etc.
3. CLASSIFY the remaining pairs and RANK them by transition orderings.

The final proposed answer is the most highly ranked candidate. An example of the algorithm in action is sketched using the following short discourse segment. Here we try to resolve the pronouns in (2.4d).

2.4a Brennan drives an Alfa Romeo.
Cb = { }; Cf = {Brennan, Alfa Romeo}
2.4b She drives too fast.
Cb = {She: Brennan}; Cf = {She: Brennan}
2.4c Friedman races her on weekends.
Cb = {her: Brennan}; Cf = {Friedman, her: Brennan, weekend}
2.4d ***She*** often beats ***her***.

---

[1] Walker (1989) does not explain the discrepancy between her evaluation and that of Hobbs. Factors that may account for this are (slight) difference in the parse tree representation and difference among the test texts.

Candidates	Cb	Cp	Transition
{She: Friedman} {her: Brennan}	Friedman	Friedman	SMOOTH-SHIFT
{She: Brennan} {her: Friedman}	Friedman	Brennan	ROUGH-SHIFT

Table 2.2 BFP example

Here we see the use of the extended SHIFT transitions. Since SMOOTH-SHIFT is preferred to ROUGH-SHIFT, the assignment <{She: Friedman} {her: Brennan}> is proposed.

As in the case of Hobbs' algorithm, there are a few implicit assumptions and assumed *knowledge* some of which are:

- implicitly assumes intersentential antecedent. Segment initial sentences are the only situation where an intrasentential antecedent is preferred
- knowledge about gender/number/animacy agreement
- knowledge of full noun phrase coreference, e.g. "Carl J. Pollard" = "Carl" = "Pollard"

Marilyn Walker(Walker 1989) reports a manual evaluation of the BFP algorithm (there is no experiment result reported in BFP 1987). The test data consists of three texts and the average result is 76.5% (average of 90%, 79%, and 60.5%).

2.3 RAP: Resolution of Anaphora Procedure

Lappin and Leass (Lappin and Leass 1994) build a system called ***Resolution of Anaphora Procedure*** (henceforth RAP). The approach derives from syntactic structures a measure of discourse salience and uses a simple dynamic model of attentional state to resolve the reference of a pronoun. A variety of intrasentential syntactic factors ("*salience factors*" as they call them) are employed. I will describe them in a little more detail not only because they help

the understanding of the algorithm but also because they show the connections between the RAP approach and our statistical approach (in Chapter 3). Most of the factors are straightforward: *sentence recency* (i.e. intrasentential or intersentential antecedents), *subject emphasis* (i.e. being a subject), *accusative emphasis* (i.e. being a direct object), and *indirect object and oblique complement emphasis*. Two other factors are *head noun emphasis* (i.e. not being contained within another noun phrase) and *non-adverbial emphasis* (i.e. not being contained in an adverbial prepositional phrase demarcated by a separator). These two factors penalize NPs in certain embedded constructions. Examples of NPs not receiving the head noun emphasis are:

> the assembly in *bay C*
> the connector on *the flat cable*

since they are embedded in another NP. Examples of NPs not receiving the non-adverbial emphasis are:

> Throughout *the first section of this guide*, these symbols are used ...
> In *the Panel definition panel*, select the "Specify" option.

These are usually the NPs occurring in a preposed prepositional phrase.

Each discourse referent has some salience factor(s) associated with it. Each salience factor, in turn, has a *weight* associated with it reflecting its relative contribution to the total salience of individual discourse referents. Initial weights of existing discourse referents are degraded by a factor of two as a new sentence is processed. This degradation, in effect, claims preference of intrasentential antecedents over intersentential ones. When the weight of a salience factor drops down to zero, the factor is removed.

The last element of this system is the use of equivalent classes. All discourse referents that are anaphorically linked form an equivalence class. As the name suggests, the coreference relation is reflexive, symmetric, and transitive. Simply put, members in an equivalence class all refer to the same discourse entity. The weight of an equivalence class is then the sum of the

current weights of all salience factors with which at least one member of the class is associated.

The RAP is now ready to proceed as follows. Upon encountering a sentence:

1. Classify each noun phrase in the sentence (definite NP, pleonastic pronoun[1], other pronoun, or indefinite NP).
2. Apply salience factors to these NPs as appropriate.
3. Apply the syntactic filter to rule out NPs on syntactic grounds.
4. Run the binding procedure for any non-pleonastic pronouns present in the sentence:
   a. A list of possible antecedents is created containing the most recent discourse referent of each equivalence class.
   b. Process each candidate. This includes calculating its salience weight, locally adjusting the weight, threshold testing, checking gender and number agreement, etc.
   c. Select the candidate with the highest salience weight. Proximity is used to resolve ties.

The syntactic filter, in essence, observes the complementary distributions of reflexive pronouns and regular pronouns (see section §1.1). Various domains are defined such as *argument* domain, *adjunct* domain, and so on. Constraints are then put on the coreference possibility within each domain. For example, a pronoun is non-coreferential with an NP if it is in the argument domain of that NP. This rules out the possibility of "*her*" coreferring with "*The woman*" in sentence (2.5):

2.5 The woman likes her.

The tests for pleonastic pronouns are partly syntactic and partly lexical. A set of adjectives and verbs that usually occur with pleonastic *IT*s are identified. In RAP, there are total fifteen adjectives such as *important*, *necessary*, etc. and eleven verbs such as *believe*, *assume*, *seem*, *appear*, etc. In addition, seven constructions are set up for the recognition of pleonastic ITs:

- It is ADJECTIVE that S

---

[1] When the pronoun It serves as a dummy subject in a cleft sentence, its use is called *pleonastic*. For example:

It is important to recognize pleonastic pronouns.

The "*It*" is used pleonastically.

- It is ADJECTIVE (for NP) to VP
- It is V-ed that S
- It seems/appears/means/follows (that) S
- NP makes/finds it ADJECTIVE (for NP) to VP
- It is time to VP
- It is thanks to NP that S

Sentences are checked against these patterns for matches.

One thing that is not clear to the author is how the gender/number/person agreement is tested. In step 5 of the algorithm, a morphologic filter, having access to a lexicon is said to be responsible for ensuring this agreement. How exactly the lexicon is structured and more importantly, how it is obtained is unclear. Extensive experiments with salience weighting are carried out on a training corpus to maximize RAP's performance. RAP achieves 85% accuracy on the training data. A test set of 345 pairs of sentences is then selected and filtered[1]. RAP achieves 86% on the test data.

2.4 Mitkov's algorithm

Mitkov (Mitkov 1998) addresses the anaphora problem by deliberately limiting the extent to which it relies on domain knowledge. The targets are anaphora in a specific genre. It is developed with the specific goal of avoiding complex syntactic, semantic, and discourse analysis. Parse trees are not used. The input is a part-of-speech (POS) tagged text. After a list of possible[1] noun phrases that precede the pronoun is identified, a set of genre-specific antecedent indicators are applied to each candidate. Similar to the salience weighting used in RAP, each candidate is assigned a score (-1, 0, 1 or 2) for each indicator. The candidate with the highest score is proposed as the antecedent. The indicators themselves are empirically determined. Each indicator more and less reflects some kind of salience. Some example indicators are:

---

[1] The test data is filtered so that it meets a few conditions. For example, for each pronoun occurrence in the set, it is made sure that the actual antecedent NP appears in the candidate list.

- *Definiteness.* Definite noun phrases are favored (scoring 0) over indefinite ones (scoring –1).
- *Givenness.* The "given" information is taken to be the first noun phrase in a non-imperative sentence. Those noun phrases are deemed good candidates and score 1.
- *Lexical reiteration.* Lexically reiterated items are highly favored. A noun phrase scores 2 if repeated within the same paragraph twice or more, 1 if repeated once and 0 if not repeated.
- *Section heading preference.* Noun phrases appearing in the heading of a section is preferred.
- *Collocation pattern preference.* Candidates with identical collocation patterns with the pronoun are given this preference.
- *Term preference.* Noun phrases that are in the terminology of the genre are preferred over those that are not.
- ⋮

It is clear from the above list that some preferences are genre-specific. But it is also worth noting that many are not.

The evaluation is manually carried out on sample texts from the genre of technical manuals containing 56 anaphoric pronouns. Mitkov reports an average success rate of 89.7%.

2.5 Conclusion

In this chapter I examined four computational approaches to the anaphora problem. As different as they may seem at first sight, there are a few common points among them in terms of the anaphora resolution factors they employ.

Although none of the algorithms emphasizes on the information encoded in the pronouns themselves, i.e. gender, number, animacy, they all check antecedents for gender/number agreement with the pronoun. Hobbs algorithm proposes an antecedent only if (among other things such as the selectional restriction) the candidate noun phrase agrees with the pronoun in these features. In example (2.4d) (see section §2.2), the BFP centering algorithm will not bind either "*She*" or "*her*" to "*weekend*" simply because it fails the animacy agreement test. As we

---

[1] Impossible candidates are those that fail the gender/number/person agreement test. It is not clear how the knowledge of gender is obtained. But since this algorithm is evaluated manually, this issue seems less critical than it would if a computer program were to be written.

have seen, RAP's morphological filter and Mitkov's gender/number agreement checker also rule out such impossible candidates.

The second common ingredient in all four algorithms is the utilization of grammatical roles, particularly those of the antecedents. Though not explicit in Hobbs' algorithm, the algorithm tends to give salience to the noun phrases in subject positions since they are among the first things that a left-to-right breadth-first search will encounter. In BFP centering, the ordering of $Cf(U_i)$ is, in fact, crucial. It is derived from the grammatical roles of each entity (e.g. subject > direct object > indirect object > others) and has direct impact on the realization of an entity in U_i as a pronoun in U_{i+1}. In particular, if a pronoun occupies a subject position[1] in U_{i+1}, Rule 1 dictates that that pronoun must realize the backward-looking center of U_{i+1}, $Cb(U_{i+1})$. This, together with Constraint 3, rules out all entities of the previous utterance but the highest-ranked one in $Cf(U_i)$. This is illustrated in the following short discourse centered around "*Tony*".

2.6a. Tony called Mike at 6am in the morning.
Cb = {Tony}; Cf = {Tony, Mike, 6am, morning}
b. **He** was furious for being woken up so early.

The "*He*" in (2.6b) is in subject position which means it must be the Cb(2.6b). Notice the ordering of the Cf(2.6a) respects the grammatical roles of each element in it. The results of two assignments to "*He*" are shown in Table 2.3.

Antecedent	Cb(2.6b)	Cp(2.6b)	Transition
He = "Tony"	Tony	Tony	CONTINUE
He = "Mike"	Mike	Mike	SMOOTH-SHIFT

Table 2.3 Ordering of $Cf(U_{i-1})$ affects $Cb(U_i)$

[1] More generally, if there is one pronoun, then it must be the Cb. The case with pronouns in the subject position is more striking and is thus chosen as an illustration example. Even more generally, it follows from Rule 1 that if there are multiple pronouns in an utterance, then one of them must be the Cb.

By the ordering preference of Rule 2, CONTINUE ≻ SMOOTH-SHIFT, "*Tony*" is chosen as the antecedent for "*He*". In some cases, this rule can be so restrictive as to not be able to find any antecedent. Consider (2.6b') following (2.6a):

2.6b' **He** was furious with Tony for being woken up so early.

Antecedent	Cb(2.6b')	Cp(2.6b')	Transition
He = "Tony"	Tony	Tony	CONTINUE*
He = "Mike"	Tony	Mike	Rule 1 Violation

Table 2.4 Restrictiveness of Rule 1

Assigning "*Mike*" to "*He*" causes a violation of Rule 1 because the backward-looking center "*Tony*" is not realized as a pronoun while the non-Cb "*Mike*" is. Having violated Rule 1, "Mike" is ruled out. "*Tony*", in fact, will be ruled out by syntactic constraint. Hence, the BFP centering algorithm (as it was originally proposed) fails to find an antecedent for "*He*" in (2.6b'). The ordering of the Cf list is so crucial that the issue of ranking forward-looking centers has become a research subject in its own right (Cote 1998, Turan 1998, Strube & Hahn 1999).

Grammatical roles are also used in RAP. Many salience emphases identified in the system correspond directly to them. Examples are *subject emphasis*, *accusative emphasis*, and so on. Salience of each grammatical role is reflected by their weights. Some example weights are show in Table 2.5.

Salience Factor	Weight
Subject emphasis	80
Accusative emphasis	50
Indirect object emphasis	40

Table 2.5 RAP's salience weighting

In Mitkov's approach, similar ideas are employed. One of his antecedent-tracking indicators (the "*Givenness*" as he calls it) gives preference to the first noun phrase in a non-imperative sentence. Given the linearity of the English language, the first noun phrase is usually the subject and thus subjects are again preferred.

The third element shared by the four algorithms (implicitly or explicitly) is the discourse salience factor. This is on the on hand, very closely related to the grammatical roles, and on the other, related to the relative distance between an antecedent and a pronoun in a discourse segment. In Hobbs' algorithm, intrasentential antecedents are preferred over intersentential ones, i.e. it prefers closer antecedents to those farther back. In BFP centering, since the targets of that study are intersentential antecedents the recency issue does not really arise. However, the measure of discourse salience is explicitly reflected in the idea of centers. The backward-looking center of an utterance is the most central, or most salient entity in that utterance. The higher an entity ranks in the Cf list, the more salient it is and thus is thought to be more likely to be pronominalized in the next utterance. As stated before, the ranking of Cf depends on the grammatical roles of its members. In English, subjects often are identified as theme, topic, given information, "aboutness" of a sentence, etc. and hence are often the most salient elements in a discourse. In RAP, besides the use of grammatical salience, the combination of equivalence class (which, in fact, is an anaphoric chain) and sentence recency constitutes a discourse model. Recall that the candidate list in RAP is constructed by finding the most recent (relative to the pronoun in

question) entity in each equivalence class. Since the weight of an equivalence class is computed by summing the weights of all salience factors that apply to its members, the more members an equivalence class has, the higher is its weight. The more frequently an entity is mentioned in the discourse, the larger the equivalence class to which it belongs. In other words, entities frequently mentioned are more salient. In terms of salience by distance, entities of previous sentence have their weights degraded by a factor of two. This, in effect, penalizes intersentential antecedents. Finally, proximity, another distance measure, is used to break ties. Similar effects can also be seen in Mitkov's algorithm. One of his antecedent-tracking indicators is *lexical reiteration* and those reiterated entities get higher scores. Sentence recency is also used and assigns a higher score to those noun phrases occurring near the pronoun.

Gender/number/animacy agreement, grammatical roles, and discourse salience are the three important factors shared by all four algorithms. Except for gender/number/animacy agreement, the algorithms differ in the extent to which they make use of the factors and the manner in which these factors are utilized.

In the case of gender/number/animacy agreement, although it is not clear in some of the algorithms, given the limited explanation in the papers, how this knowledge is obtained, it is clear that all four algorithms have a perfect source of such information (e.g. by human judgement) and they all make full use of it. In the case of grammatical roles and discourse salience factors, the algorithms clearly differ.

Hobbs algorithm does not consider the grammatical roles of the pronouns, only those of the antecedents. In BFP centering, the grammatical role of a pronoun comes into play when there are multiple pronouns in an utterance and one them is the preferred center Cp^1 (i.e. in the subject position). By ways of Constraints and Rankings, the BFP algorithm uses the grammatical and discourse factors extensively. RAP and Mitkov's algorithms look at the grammatical roles of

pronouns only in one specific situation, that of the parallel cases. In RAP, if a candidate antecedent fills the same grammatical slot as the pronoun, its weight is increased, i.e. parallelism of grammatical roles is rewarded. In Mitkov's algorithm, the indicator called *collocation pattern* does essentially the same thing, rewarding the candidates with identical collocation patterns with the pronoun. In terms of factor utilization, Mitkov's approach is very similar to that of the RAP system. It is a much simplified version of RAP without the syntactic filter and with genre-specific knowledge.

The fact that all four seemingly different algorithms share these three common elements is no accident. We will come back to these factors in Chapter 4 after we present our statistical approach in Chapter 3. In the next chapter we will see how these factors are used in a statistical framework that achieves accuracy higher than any of the approaches described in this chapter.

---

[1] This is because Cp affects the transition type, and hence the preference among different assignments. Consider the example (2.4d) where there are two pronouns "She" and "her" and the Cp is a pronoun ("She"). In Table 2.2, we see one assignment results in SMOOTH-SHIFT and the other ROUGH-SHIFT.

Chapter 3 A Statistical Approach

In this chapter I present a computational approach to anaphora resolution within a statistical framework. Statistical approach has revolutionized natural language processing and Artificial Intelligence in general in the past few decades. It has shown remarkable success in tagging, parsing, speech recognition, word sense disambiguation, and may other areas in NLP. That is the major motivation for approaching the anaphora problem in this line of research. Surprisingly or not, it has once again demonstrated its potential.

There are two models that we experimented with. Although the second model gives the best result, I will present them both. There are two reasons to this. The first one is that although the second model is in a way an extension to the first model, there is one component present in the first but absent in the second, which will lead to some interesting discussions. The other reason is that comparisons of the two models require an understanding of both of them.

In section 3.1, I will present the two models. In section 3.2, each component of the models is explained. I will then discuss some implementation issues in section 3.3. The experimental results can be found in section 3.4. I will conclude this chapter with a comparison of our model and the four previous approaches discussed in Chapter 2.

3.1 Two probabilistic models

3.1.1 The basic model

We treat the antecedent of a pronoun as a ***random variable A(ρ)*** where ρ denotes the pronoun in question. Given the context in the discourse surrounding the pronoun ρ, we compute the ***probability*** that some noun phrase is the antecedent of ρ. In other words, we want to compute $P(A(\rho) = a \mid context)$ where **a** is the candidate antecedent under consideration. This probability is computed for every member in a candidate list (the gathering of which will be explained shortly).

The proposed antecedent is the one that maximizes this probability. In other words, we want to assign a referent to the pronoun ρ which is the most likely antecedent in a given context. The context in the conditioning events is derived from various sources of linguistic information.

To put this more formally, let $F(\rho)$ denote a function from pronoun ρ to its antecedent, then:

$$F(\rho) = \arg\max_{a} P(A(\rho) = a \mid context)$$
$$= \arg\max_{a} P(A(\rho) = a \mid \rho, \vec{d}_H, \vec{W}, h, t, l, \vec{M}, S_\rho, f_\rho) \qquad (3.1)$$

where

- $A(\rho)$ is a random variable denoting the antecedent of the pronoun ρ
- a is a proposed antecedent
- $\overrightarrow{d_H}$ is a vector quantity specifying the Hobbs distance of each candidate from ρ. It is obtained by running Hobbs' algorithm. The first antecedent Hobbs' algorithm proposes is at distance 1 ($d_H = 1$), the second is at distance 2 ($d_H = 2$), and so on.
- $\overrightarrow{W}$ is the list of candidate antecedents to be considered. It is also a vector quantity and in our experiment we consider 25 candidates for every pronoun
- h is the lexical item governing ρ. It is usually a verb. For example, the head of "*he*" in "*he said ...*" is the verb "*said*"
- t is the type of phrase of the proposed antecedent. It will always be a noun phrase (NP)
- l is the category label of the maximal projection of the governor h. In the above example of "*he said...*", the type of the head would be S. One other typical situation is phrases like "*eat it*" where the head is the verb "*eat*" and its type would be VP.
- $\overrightarrow{M}$ is the number of times each candidate is mentioned up to that point in the discourse and is a vector quantity

- S_ρ is the sentence number of the sentence in which ρ finds itself. Sentences in a discourse segment are numbered sequentially.
- f_ρ is the form of the pronoun ρ and is one of the following three: reflexive, possessive, or regular.

When viewed in this way, a can be regarded as an index into the vectors ($\overline{d_H}$, $\overline{W}$, $\overline{M}$) that specifies which value is relevant to the particular choice of antecedent. Needless to say, the probabilities of equation (3.1) are too specific to have any hope of obtaining, or even guessing. We therefore must make simplifying independence assumptions and decompose this equation into statistically manageable components. The decomposition makes use of **Bayes Theorem**. In the following derivation, n is the number of candidates. The independence assumptions and explanations of each step are described at the end of the derivation[1].

$$P(A(\rho)=a \mid \rho,\vec{d}_{H.}\bar{W},h,t,l,\bar{M},S_{\rho.}f_\rho) \tag{3.2}$$

$$=P(a \mid \rho,\vec{d}_{H.}\bar{W},h,t,l,\bar{M},S_{\rho.}f_\rho) \tag{3.3}$$

$$=\frac{P(a,\rho,\vec{d}_{H.}\bar{W},h,t,l,\bar{M},S_{\rho.}f_\rho)}{P(\rho,\vec{d}_{H.}\bar{W},h,t,l,\bar{M},S_{\rho.}f_\rho)} \tag{3.4}$$

$$=\frac{P(a,\rho,\vec{d}_{H.}\bar{W},h,t,l,f_\rho \mid \bar{M},S_\rho)P(\bar{M},S_\rho)}{P(\rho,\vec{d}_{H.}\bar{W},h,t,l,f_\rho \mid \bar{M},S_\rho)P(\bar{M},S_\rho)} \tag{3.5}$$

$$=\frac{P(a,\rho,\vec{d}_{H.}\bar{W},h,t,l,f_\rho \mid \bar{M},S_\rho)}{P(\rho,\vec{d}_{H.}\bar{W},h,t,l,f_\rho \mid \bar{M},S_\rho)} \tag{3.6}$$

$$=\frac{P(a \mid \bar{M},S_\rho)P(\rho,\vec{d}_{H.}\bar{W},h,t,l,f_\rho \mid a,\bar{M},S_\rho)}{P(\rho,\vec{d}_{H.}\bar{W},h,t,l,f_\rho \mid \bar{M},S_\rho)} \tag{3.7}$$

$$\approx\frac{P(a \mid \bar{M},S_\rho)P(\vec{d}_H \mid a,f_\rho)P(\rho,\bar{W},h,t,l,f_\rho \mid a,\bar{M},S_\rho)}{P(\vec{d}_H)P(\rho,\bar{W},h,t,l,f_\rho \mid \bar{M},S_\rho)} \tag{3.8}$$

---

[1] The "≈" means the step makes use of an independence assumption.

$$\approx \frac{P(a|\vec{M},S_\rho)P(\vec{d}_H|a,f_\rho)P(h,t,l)P(\rho,\vec{W},f_\rho|a,\vec{M},S_\rho,h,t,l)}{P(\vec{d}_H)P(h,t,l)P(\rho,\vec{W},f_\rho|\vec{M},S_\rho,h,t,l)} \tag{3.9}$$

$$= \frac{P(a|\vec{M},S_\rho)P(\vec{d}_H|a,f_\rho)P(\vec{W}|a,\vec{M},S_\rho,h,t,l)P(\rho,f_\rho|a,\vec{M},S_\rho,h,t,l,\vec{W})}{P(\vec{d}_H)P(\vec{W}|\vec{M},S_\rho,h,t,l)P(\rho,f_\rho|\vec{M},S_\rho,h,t,l,\vec{W})} \tag{3.10}$$

$$\approx \frac{P(a|\vec{M},S_\rho)P(\vec{d}_H|a,f_\rho)P(\vec{W}|a,h,t,l)P(\rho|a,\vec{M},S_\rho,h,t,l,\vec{W})P(f_\rho|a,\vec{M},S_\rho,h,t,l,\vec{W})}{P(\vec{d}_H)P(\vec{W}|t)P(\rho|\vec{M},S_\rho,h,t,l,\vec{W})P(f_\rho|\vec{M},S_\rho,h,t,l,\vec{W})} \tag{3.11}$$

$$\approx \frac{P(a|\vec{M},S_\rho)P(\vec{d}_H|a,f_\rho)P(\vec{W}|a,h,t,l)P(\rho|a,\vec{M},S_\rho,h,t,l,\vec{W})P(f_\rho|a,\vec{M},S_\rho,h,t,l,\vec{W})}{P(\vec{d}_H)P(\vec{W}|t)P(\rho|\vec{M},S_\rho,h,t,l,\vec{W})P(f_\rho|\vec{M},S_\rho,h,t,l,\vec{W})} \tag{3.12}$$

$$= \frac{P(a|\vec{M},S_\rho)P(\vec{d}_H|a,f_\rho)P(\vec{W}|a,h,t,l)P(\rho|a,\vec{W})}{P(\vec{d}_H)P(\vec{W}|t)P(\rho)} \tag{3.13}$$

$$\approx \frac{P(a|M_a,S_\rho)[\prod_{i=1}^{n}P(d_i|a,f_\rho)][\prod_{i=1}^{n}P(W_i|a,h,t,l)]\,P(\rho|a,W_a)}{[\prod_{i=1}^{n}P(d_i)][\prod_{i=1}^{n}P(W_i|t)]\,P(\rho)} \tag{3.14}$$

$$\approx \frac{P(a|M_a,S_\rho)P(d_a|a,f_\rho)P(W_a|a,h,t,l)P(\rho|a,W_a)}{P(d)P(W_a|t)P(\rho)} \tag{3.15}$$

$$= \frac{P(a|M_a,S_\rho)P(d_a|a,f_\rho)P(W_a|a,h,t,l)P(\rho|a,W_a)}{P(W_a|t)} * \frac{1}{P(d)P(\rho)} \tag{3.16}$$

$$\propto P(d_a|a,f_\rho) * P(\rho|a,W_a) * \frac{P(W_a|a,h,t,l)}{P(W_a|t)} * P(a|M_a,S_\rho) \tag{3.17}$$

Using statistics conventions, the probability of a random variable X taking a particular value x, P(X=x) is usually written simply as P(x). Equation (3.3) is the result of using this convention. Equations (3.4) to (3.13) is a series of applications of Bayes' formula. During these steps, the following independence relations are applied:

- Hobbs' distances are independent of the mention counts M, the sentence position of the pronoun S_ρ, the pronoun ρ, the head environment surrounding the pronoun (h,t,l), or the words in the candidates W. Distances depend only on the choice of the antecedent a and the form of the pronoun f_ρ:

$$P(\vec{d}_H \mid a, \vec{M}, S_\rho, \rho, \vec{W}, h, t, l, f_\rho) = P(\vec{d}_H \mid a, f_\rho)$$

  This relation is applied to both the numerator and the denominator of equation (3.7). In the denominator, since the antecedent a is not assumed, the distances are independent of f_ρ.

- The head information, consisting of the head of the pronoun (h), the type of that head (l), and the type of the candidate antecedent phrase (t), is independent of other contextual parameters such as the choice of the antecedent a, the pronoun ρ, the mention counts M, etc.

$$P(h,t,l \mid a, \vec{M}, S_\rho, \rho, \vec{W}, f_\rho) = P(h,t,l)$$

  This relation is applied to equation (3.8), resulting in equation (3.9).

- The words of the candidates depend only on the head information and the choice of the antecedent, and are independent of everything else:

$$P(\vec{W} \mid a, \vec{M}, S_\rho, h, t, l) = P(\vec{W} \mid a, h, t, l)$$

Without a particular antecedent, the words depend only on their own type, t:

$$P(\vec{W} \mid \vec{M}, S_\rho, h, t, l) = P(\vec{W} \mid t)$$

Also notice that ρ and its form f_ρ are independent. These two observations are applied to (3.10), yielding (3.11).

- The next independence relation is that the choice of pronoun depends only on the words in the candidates, i.e.

$$P(\rho \mid a, \vec{M}, S_\rho, h, t, l, \vec{W}) = P(\rho \mid a, \vec{W})$$

This is applied to equation (3.11). Note that in the denominator, the index a is not given. Thus ρ and W are independent of each other because no particular antecedent is assumed.

- The form of the pronoun f_ρ is independent of all other parameters in (3.11), i.e.

$$P(f_\rho \mid a, \vec{M}, S_\rho, h, t, l, \vec{W}) = P(f_\rho)$$

From equation (3.13) on, further independence assumptions are made. They are:

- The probability that a given noun phrase is the correct antecedent depends only on the mention counts of itself and is independent of the mention counts of other noun phrases. Thus

$$\begin{aligned} P(a \mid \vec{M}, S_\rho) &= P(a \mid M_1, M_2, \cdots, M_a, \ldots M_n, S_\rho) \\ &= P(a \mid M_a, S_\rho) \end{aligned}$$

- The distances in d_H are independent of each other, i.e. $P(d_{i+1} \mid d_i) = P(d_{i+1})$. Here d_i is the i^{th} element of the vector d_H. To make the notation more readable, we write d_i instead of $(d_H)_i$. Thus

$$P(\vec{d}_H \mid a, f_\rho) = \prod_{i=1}^{n} P(d_i \mid a, f_\rho)$$

and

$$P(\vec{d}_H) = \prod_{i=1}^{n} P(d_i)$$

- It is reasonable to assume that the candidates in $\overline{\mathrm{W}}$ are independent of each other. In other words $P(W_{i+1} \mid W_i, h, t, l, a) = P(W_{i+1} \mid h, t, l, a)$. Thus

$$P(\vec{W} \mid t) = \prod_{i=1}^{n} P(W_i \mid t)$$

and

$$P(\vec{W} \mid a,h,t,l) = \prod_{i=1}^{n} P(W_i \mid a,h,t,l)$$

- If we treat a as an index into the vector $\overline{\mathrm{W}}$, then (a, $\overline{\mathrm{W}}$) is simply the a^{th} element in the list. We assume that the selection of the pronoun is independent of the candidates other than the antecedent. Hence

$$\begin{aligned} P(\rho \mid a,\vec{W}) &= P(\rho \mid a,W_1,W_2,\ldots,W_a,\ldots,W_n) \\ &= P(\rho \mid a,W_a) \end{aligned}$$

Application of the above brings (3.14) from (3.13). In the denominator (3.14), note that the prior distribution on distance is uniform which enables us to drop the subscript and simply put $P(d_i) = P(d)$. Many terms in (3.14) cancel each other in the following way:

$$\begin{aligned} \frac{\prod_{i=1}^{n} P(d_i \mid a, f_\rho)}{\prod_{i=1}^{n} P(d_i)} &= \frac{P(d_1)P(d_2)\cdots P(d_a \mid a, f_\rho)\cdots P(d_n)}{P(d)P(d)\cdots P(d)\cdots P(d)} \\ &= \frac{P(d_a \mid a, f_\rho)}{P(d)} \end{aligned}$$

This is because

$$P(d_i \mid a, f_\rho) = \begin{cases} P(d_i) & \text{if } i \neq a \\ P(d_a \mid a, f_\rho) & \text{if } i = a \end{cases}$$

Another cancellation is the following

$$\frac{\prod_{i=1}^{n} P(W_i \mid a,h,t,l)}{\prod_{i=1}^{n} P(W_i \mid t)} = \frac{P(W_1 \mid t)P(W_2 \mid t)\cdots P(W_a \mid a,h,t,l)\cdots P(W_n \mid t)}{P(W_1 \mid t)P(W_2 \mid t)\cdots P(W_a \mid t)\cdots P(W_n \mid t)}$$

$$= \frac{P(W_a \mid a,h,t,l)}{P(W_a \mid t)}$$

And this is because

$$P(W_i \mid a,h,t,l) = \begin{cases} P(W_i \mid t) & \text{if } i \neq a \\ P(W_a \mid a,h,t,l) & \text{if } i = a \end{cases}$$

After these cancellations, equation (3.14) becomes (3.15). Equation (3.16) is a rearrangement of (3.14). In (3.16) the term

$$\frac{1}{P(d)P(\rho)}$$

is the same for every candidate antecedent and thus can be eliminated, resulting in equation (3.17) proportional to the original equation (3.2). Returning to F(ρ) in equation (3.1), we see that maximizing (3.1) is equivalent to maximizing (3.17). Thus

$$F(\rho) = \arg\max_a P(A(\rho) = a \mid \rho, \vec{d}_H, \vec{W}, h,t,l, \vec{M}, S_\rho, f_\rho) \qquad (3.18)$$

$$= \arg\max_a P(d_a \mid a, f_\rho) P(\rho \mid a, W_a) \frac{P(W_a \mid a,h,t,l)}{P(W_a \mid t)} P(a \mid M_a, S_\rho) \qquad (3.19)$$

The components of equation (3.19) are not only statistically manageable, but have very intuitive meanings as well. I will briefly explain them here and will examine them in detail in the next section (§3.2).

The first component $P(d_a \mid a, f_p)$ computes the probability of an antecedent occurring at a particular Hobbs' distance. Hobbs distances are obtained by running Hobbs' algorithm (the section on implementation §3.3 will have all the specifics). As described in section §2.1, Hobbs' algorithm is a syntactic approach and traverses the parse trees in a particular order making sure to rule out those noun phrases that violate the coreferential binding principles. By using this algorithm we are, in effect, taking syntactic constraints into account. The probability $P(p \mid a, W_a)$ encodes the gender/number/animacy information since it computes the probability of choosing a particular pronoun (HE, SHE, IT, etc.) given the antecedent word. Lexical semantics are reflected in $\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$. It asks for the probability of a candidate being the correct antecedent given the head of the pronoun, i.e. how "likely" can the candidate (W_a) be expected to be in the environment created by the head h. Using the canonical example of the verb "*eat*", this statistic measures the plausibility of a noun being the direct object of "*eat*" if the pronoun follows "*eat*", i.e. $P(W_a \mid a, \text{"eat"}, NP, VP)$. One hopes that this statistic would pick out objects that are foods. If the pronoun precedes "*eat*", then this statistic would compute the probability of an object performing the act of "*eat*", i.e. $P(W_a \mid a, \text{"eat"}, NP, S)$ in which case one would like to see an animate object being selected. The last term $P(a \mid M_a, S_p)$ approximates discourse topics. The higher the candidate's mention counts (M_a), the more likely it is the topic of the story. As topics are more salient than other entities in a story, higher mention counts indicate that the candidate is more likely to be the correct antecedent. The position of a pronoun in a discourse can have influence on the mention count of its referent. In other words, the nearer the end of the discourse segment a pronoun occurs, the more probable it is that its antecedent has been mentioned several

times. Hence the S_ρ is in the conditioning event. Table 3.1 summarizes the meanings of these components.

Statistic	Factor Represented
$P(d_a \mid a, f_\rho)$	Syntax
$P(\rho \mid a, W_a)$	Gender/Number/Animacy
$\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$	Lexical semantics
$P(a \mid M_a, S_\rho)$	Discourse topic/salience

Table 3.1 Factors in the Basic model

3.1.2 The syntactic-prominence model

The second model is built on top of the previous model and is an extension of it (with an exception). It is set in the same statistical framework but is different in that it explores a different set of contextual parameters. Specifically, the context is a superset of that of the previous one, i.e. more information is incorporated. The additional information comes from the grammatical roles of the candidates and the pronoun, and sentence recency measures of the candidates. Using formal notations, this model maximizes the following:

$$F(\rho) = \arg\max_a P(A(\rho) = a \mid \rho, \vec{d}_H, \vec{W}, \vec{M}, S_\rho, f_\rho, G_\rho, \vec{G}_w, \vec{S}_w) \qquad (3.20)$$

where the additional contexts are:

- G_ρ is the grammatical role of the pronoun ρ
- $\overrightarrow{G_w}$ is a vector containing the grammatical roles of each candidate in the $\overrightarrow{W}$ list

- $\overline{S_w}$ is a vector containing the sentence numbers of each candidate relative to the one in which the pronoun occurs. Intrasentential candidates occur at relative sentence position 0, candidates from the immediate preceding sentence is at sentence position 1, and so on.

For the same reason as before (section §3.1.1) this equation needs to be decomposed, i.e. factorized so we can compute the probability. The derivation makes use of Bayes' inversion formula together with a set of independence assumptions which are discussed following the derivations.

$$P(A(\rho)=a \mid \rho,\vec{d}_H,\vec{W},\vec{M},S_\rho,G_\rho,\vec{G}_w,f_\rho,\vec{S}_w) \tag{3.21}$$

$$=P(a \mid \rho,\vec{d}_H,\vec{W},\vec{M},S_\rho,G_\rho,\vec{G}_w,f_\rho,\vec{S}_w) \tag{3.22}$$

$$=\frac{P(a,\rho,\vec{d}_H,\vec{W},\vec{M},S_\rho,G_\rho,\vec{G}_w,f_\rho,\vec{S}_w)}{P(\rho,\vec{d}_H,\vec{W},\vec{M},S_\rho,G_\rho,\vec{G}_w,f_\rho,\vec{S}_w)} \tag{3.23}$$

$$\approx\frac{P(\rho \mid a,\vec{W})P(\vec{d}_H \mid a,f_\rho)P(\vec{W})P(\vec{G}_w,\vec{S}_w,\vec{M},S_\rho,G_\rho \mid a,f_\rho)P(a,f_\rho)}{P(\rho \mid \vec{W})P(\vec{d}_H \mid f_\rho)P(\vec{W})P(\vec{G}_w,\vec{S}_w,\vec{M},S_\rho,G_\rho \mid f_\rho)P(f_\rho)} \tag{3.24}$$

$$\approx\frac{P(\rho \mid a,\vec{W})P(\vec{d}_H \mid a,f_\rho)P(\vec{G}_w \mid a,G_\rho)P(\vec{S}_w,\vec{M},S_\rho,G_\rho \mid a,f_\rho)P(a)P(f_\rho)}{P(\rho)P(\vec{d}_H)P(\vec{G}_w)P(\vec{S}_w,\vec{M},S_\rho,G_\rho \mid f_\rho)P(f_\rho)} \tag{3.25}$$

$$=\frac{P(\rho \mid a,\vec{W})P(\vec{d}_H \mid a,f_\rho)P(\vec{G}_w \mid a,G_\rho)P(\vec{S}_w \mid a,f_\rho)P(\vec{M},S_\rho,G_\rho \mid a,f_\rho)P(a)}{P(\rho)P(\vec{d}_H)P(\vec{G}_w)P(\vec{S}_w)P(\vec{M},S_\rho,G_\rho \mid f_\rho)} \tag{3.26}$$

$$\approx\frac{P(\rho \mid a,\vec{W})P(\vec{d}_H \mid a,f_\rho)P(\vec{G}_w \mid a,G_\rho)P(\vec{S}_w \mid a,f_\rho)P(\vec{M},S_\rho \mid a,f_\rho,G_\rho)P(G_\rho)P(a)}{P(\rho)P(\vec{d}_H)P(\vec{G}_w)P(\vec{S}_w)P(\vec{M},S_\rho \mid f_\rho,G_\rho)P(G_\rho)} \tag{3.27}$$

$$=\frac{P(\rho|a,\vec{W})P(\vec{d}_H|a,f_\rho)P(\vec{G}_w|a,G_\rho)P(\vec{S}_w|a,f_\rho)P(\vec{M}|a,f_\rho,G_\rho,S_\rho)P(S_\rho|a,f_\rho,G_\rho)P(a)}{P(\rho)P(\vec{d}_H)P(\vec{G}_w)P(\vec{S}_w)P(\vec{M}|f_\rho,G_\rho,S_\rho)P(S_\rho|f_\rho,G_\rho)} \quad (3.28)$$

$$=\frac{P(\rho|a,\vec{W})P(\vec{d}_H|a,f_\rho)P(\vec{G}_w|a,G_\rho)P(\vec{S}_w|a,f_\rho)P(\vec{M}|a,G_\rho,S_\rho)P(S_\rho)P(a)}{P(\rho)P(\vec{d}_H)P(\vec{G}_w)P(\vec{S}_w)P(\vec{M})P(S_\rho)} \quad (3.29)$$

$$=\frac{P(\rho|a,\vec{W})P(\vec{d}_H|a,f_\rho)P(\vec{G}_w|a,G_\rho)P(\vec{S}_w|a,f_\rho)P(\vec{M}|a,G_\rho,S_\rho)P(a)}{P(\rho)P(\vec{d}_H)P(\vec{G}_w)P(\vec{S}_w)P(\vec{M})} \quad (3.30)$$

$$=\frac{P(\rho|a,W_a)[\prod_{i=1}^{n}P(d_i|a,f_\rho)][\prod_{i=1}^{n}P(G_{w_i}|a,G_\rho)][\prod_{i=1}^{n}P(S_{w_i}|a,f_\rho)][\prod_{i=1}^{n}P(M_i|a,G_\rho,S_\rho)]P(a)}{P(\rho)[\prod_{i=1}^{n}P(d_i)][\prod_{i=1}^{n}P(G_{w_i})][\prod_{i=1}^{n}P(S_{w_i})][\prod_{i=1}^{n}P(M_i)]} \quad (3.31)$$

$$=\frac{P(\rho|a,W_a)P(d_a|a,f_\rho)P(G_{w_a}|a,G_\rho)P(S_{w_a}|a,f_\rho)P(M_a|a,G_\rho,S_\rho)P(a)}{P(\rho)P(d)P(G_w)P(S_w)P(M_a)} \quad (3.32)$$

$$=\frac{P(\rho|a,W_a)P(d_a|a,f_\rho)P(G_{w_a}|a,G_\rho)P(S_{w_a}|a,f_\rho)P(M_a|a,G_\rho,S_\rho)}{P(M_a)}$$

$$*\frac{P(a)}{P(\rho)P(d)P(G_w)P(S_w)} \quad (3.33)$$

$$\propto P(\rho|a,W_a)*P(d_a|a,f_\rho)*P(G_{w_a}|a,G_\rho)*P(S_{w_a}|a,f_\rho)*\frac{P(M_a|a,G_\rho,S_\rho)}{P(M_a)} \quad (3.34)$$

Equation (3.23) is the result of directly applying Bayes' formula to (3.22). In (3.23) we observe that in this context:

- The choice of pronoun depends only on the candidates:

$$P(\rho \mid a, \vec{d}_H, \vec{W}, \vec{M}, S_\rho, G_\rho, \vec{G}_w, f_\rho, \vec{S}_w) = P(\rho \mid a, \vec{W})$$

- Hobbs' distances may be influenced by the form of the pronoun but not anything else[1]:

$$P(\vec{d}_H \mid a, \vec{W}, \vec{M}, S_\rho, G_\rho, \vec{G}_w, f_\rho, \vec{S}_w) = P(\vec{d}_H \mid a, f_\rho)$$

- The phrases of the candidates do not depend on their grammatical roles, their mention counts, their relative sentence positions, the grammatical role of the pronoun, the sentence number of the pronoun, or the form of the pronoun:

$$P(\vec{W} \mid \vec{G}_w, \vec{M}, \vec{S}_w, G_\rho, S_\rho, f_\rho, a) = P(\vec{W})$$

These relations are applied to both the numerator and the denominator of equation (3.23) which results in (3.24). In the denominator of (3.24), there is not a particular choice of a candidate (i.e. the index a is missing) and hence ρ and d_H are independent of their respective conditioning events. This is shown in the denominator of equation (3.25). In (3.24), we further observe that:

- The grammatical role of the antecedent is closely related to the grammatical role of the pronoun, but not anything else in the remaining parameters:

$$P(\vec{G}_w \mid \vec{S}_w, \vec{M}, S_\rho, G_\rho, a, f_\rho) = P(\vec{G}_w \mid a, G_\rho)$$

This brings equation (3.25) in which the relative sentence numbers of the candidates are independent of the rest of the parameters:

$$P(\vec{S}_w, \vec{M}, S_\rho, G_\rho \mid a, f_\rho) = P(\vec{S}_w \mid a, f_\rho) P(\vec{M}, S_\rho, G_\rho \mid a, f_\rho)$$

This is reflected in equation (3.26). Applying Bayes' inversion to $P(\overline{M}, S_\rho, G_\rho)$ yields (3.27) and then (3.28). Two independence relations are applied to (3.28):

- The mention counts of candidate antecedents are independent of the form of the pronoun:

$$P(\vec{M} \mid a, f_\rho, S_\rho, G_\rho) = P(\vec{M} \mid a, S_\rho, G_\rho)$$

- The sentence number in which the pronoun occurs is independent of either its form or its grammatical role:

$$P(S_\rho \mid a, f_\rho, G_\rho) = P(S_\rho)$$

This results in equation (3.29). Cancellations of common terms in the numerator and the denominator yield (3.30). The vector expansions assume the same pairwise independence relations described in the last section (§3.1.1). This is shown in equation (3.31). The denominator of that equation is essentially a product of various prior distributions:

$$P(G_{w_i} \mid a, G_\rho) = \begin{cases} P(G_{w_a} \mid a, G_\rho) & \text{if i} = \text{a} \\ P(G_{w_i}) & \text{if i} \neq \text{a} \end{cases}$$

$$P(S_{w_i} \mid a, f_\rho) = \begin{cases} P(S_{w_a} \mid a, f_\rho) & \text{if i} = \text{a} \\ P(S_{w_i}) & \text{if i} \neq \text{a} \end{cases}$$

$$P(M_i \mid a, G_\rho, S_\rho) = \begin{cases} P(M_a \mid a, G_\rho, S_\rho) & \text{if i} = \text{a} \\ P(M_i) & \text{if i} \neq \text{a} \end{cases}$$

This then leads to:

---

[1] The distances are also dependent on the grammatical role of the pronoun G_ρ. However empirical results show that the extra conditioning on G_ρ, $P(d_H \mid a, f_\rho, G_\rho)$ does not improve the overall performance and that conditioning on f_ρ alone works slightly better than conditioning on G_ρ alone.

$$\frac{\prod_{i=1}^{n} P(G_{w_i} \mid a, G_\rho)}{\prod_{i=1}^{n} P(G_{w_i})} = \frac{P(G_{w_1})P(G_{w_2}) \cdots P(G_{w_a} \mid a, G_\rho) \cdots P(G_{w_n})}{P(G_{w_1})P(G_{w_2}) \cdots P(G_{w_a}) \cdots P(G_{w_n})}$$

$$= \frac{P(G_{w_a} \mid a, G_\rho)}{P(G_{w_a})}$$

$$= \frac{P(G_{w_a} \mid a, G_\rho)}{P(G_w)}$$

$$\frac{\prod_{i=1}^{n} P(S_{w_i} \mid a, f_\rho)}{\prod_{i=1}^{n} P(S_{w_i})} = \frac{P(S_{w_1})P(S_{w_2}) \cdots P(S_{w_a} \mid a, f_\rho) \cdots P(S_{w_n})}{P(S_{w_1})P(S_{w_2}) \cdots P(S_{w_a}) \cdots P(S_{w_n})}$$

$$= \frac{P(S_{w_a} \mid a, f_\rho)}{P(S_{w_a})}$$

$$= \frac{P(S_{w_a} \mid a, f_\rho)}{P(S_w)}$$

and

$$\frac{\prod_{i=1}^{n} P(M_i \mid a, G_\rho, S_\rho)}{\prod_{i=1}^{n} P(M_i)} = \frac{P(M_1)P(M_2) \cdots P(M_a \mid a, G_\rho, S_\rho) \cdots P(M_n)}{P(M_1)P(M_2) \cdots P(M_a) \cdots P(M_n)}$$

$$= \frac{P(M_a \mid a, G_\rho, S_\rho)}{P(M_a)}$$

As we did in the last section, without a particular choice of antecedent, the prior distributions can be assumed to be uniform and hence the subscripts in the denominator are dropped. One exception is that of the mention counts because here the distribution is clearly not uniform. After all the cancellations shown above, we arrive at equation (3.32). (3.33) is a rearrangement of (3.32) in which the prior terms are grouped together. Since the priors are uniform, (3.34) is

proportional to (3.33). Maximizing the original equation (3.20) is then equivalent to maximizing (3.34):

$$F(\rho) = \underset{a}{\operatorname{argmax}} P(A(\rho) = a \mid \rho, \vec{d}_H, \vec{W}, \vec{M}, S_\rho, f_\rho, G_\rho, \vec{G}_w, \vec{S}_w) \tag{3.35}$$

$$= \underset{a}{\operatorname{argmax}} P(\rho \mid a, W_a) P(d_a \mid a, f_\rho) P(G_{w_a} \mid a, G_\rho) P(S_{w_a} \mid a, f_\rho) \frac{P(M_a \mid a, G_\rho, S_\rho)}{P(M_a)} \tag{3.36}$$

Like the basic model, components in (3.36) correspond directly to linguistic factors. The $P(G_{Wa} \mid a, G_\rho)$ relates the grammatical role of the antecedent to that of the pronoun and $P(S_{Wa} \mid a, f_\rho)$ computes the sentence recency probability depending on the form of the pronoun. The factors used in this model are summarized in Table 3.2.

Statistic	Factor represented
$P(d_a \mid a, f_\rho)$	Syntax
$P(\rho \mid a, w_a)$	Gender/Number/Animacy
$P(G_w \mid a, G_\rho)$	Syntactic prominence
$P(S_w \mid a, f_\rho)$	Sentence recency
$\frac{P(M_a \mid a, S_\rho, G_\rho)}{P(M_a)}$	Discourse topic

Table 3.2 Factors in the Syntactic-prominence model

3.1.3 A special case

Both the basic model and the syntactic-prominence model are designed to handle all occurrences of anaphoric pronouns regardless of their gender class (HE/SHE/IT etc.) or number class (singular/plural). However, as described in section §2.3, the pronoun ***IT*** can sometimes act

pleonastically[1]. We would like to be able to recognize such usage of ***IT*** so that we don't falsely assign them a referent.

3.1.3.1 The equation

This task is also accomplishable in the statistical models we set up in the previous two sections. The way we approach this phenomenon is to identify a set of sentence patterns in which it is very likely for an "***it***" to behave pleonastically. We then add a "***pattern***" parameter into the contexts. The derivations are very similar to the ones presented before except that a new parameter is added. To simplify the matter, the "*pattern*" parameter is independent of the contexts already in the models. In the basic model, this addition results in:

$$F(\rho) = \arg\max_a P(A(\rho) = a \mid \rho, \vec{d}_H, \vec{W}, h, t, l, \vec{M}, S_\rho, f_\rho, pattern) \quad (3.37)$$

$$= \arg\max_a P(d_a \mid a, f_\rho) P(\rho \mid a, W_a) \frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)} P(a \mid M_a, S_\rho) P(pattern \mid a) \quad (3.38)$$

We can think of this situation as a pleonastic "*it*" having *NULL* as its referent. If the pronoun "*it*" is indeed pleonastic, equation (3.38) is then: (*pleo* is short for *pleonastic*)

$$P(d \mid pleo, f_\rho) P(\rho \mid pleo, NULL) \frac{P(NULL \mid pleo, h, t, l)}{P(NULL \mid t)} P(pleo \mid M_a, S_\rho) P(pattern \mid pleo) \quad (3.39)$$

For pleonastic ITs, the Hobbs' distance measure $P(d \mid pleo, f_\rho)$ and the lexical semantics of the antecedent (which is NULL) $\frac{P(NULL \mid pleo, h, t, l)}{P(NULL \mid t)}$ do not apply any more. Among all the pronouns, only *IT*s can be used pleonastically. Therefore $\rho \equiv IT$ and

---

[1] There are cases where "it" is used neither anaphorically nor pleonastically. Some of these are conventional unspecified referents as in

It is raining.

and the "*do it*" anaphora as in

*John wanted to jump off the cliff and Bill told him not to **do it**.*

$$P(\rho \mid pleo, NULL) = \begin{cases} 1 & if\ \rho = IT \\ 0 & otherwise \end{cases}$$

Whether or not an "*it*" is pleonastic has nothing to do with mention counts of the candidates (since there aren't any referents) or with the sentence number in which the "*it*" occurs. Thus $P(pleo \mid M_a, S_\rho) = P(pleo)$. Under these conditions (3.39) becomes simply

$$P(pleonastic)P(pattern \mid pleonastic) \qquad (3.40)$$

After adding the "*pattern*" parameter to the syntactic-prominence model, equation (3.35) becomes (in the equations below, "*ptn*" is short for "*pattern*")

$$F(\rho) = \arg\max_a P(A(\rho) = a \mid \rho, \vec{d}_H, \vec{W}, \vec{M}, S_\rho, f_\rho, G_\rho, \vec{G}_w, \vec{S}_w, ptn) \qquad (3.41)$$

$$= \arg\max_a P(\rho \mid a, W_a) P(d_a \mid a, f_\rho) P(G_{w_a} \mid a, G_\rho) P(S_{w_a} \mid a, f_\rho) \frac{P(M_a \mid a, G_\rho, S_\rho)}{P(M_a)} P(ptn \mid a) P(a) \qquad (3.42)$$

The same reasoning as that above is applied to (3.42) and we have again arrived at (3.40) for identifying pleonastic *IT*s. The *P(pleonastic)* term in equation (3.40) is simply the prior probability of an "*it*" pronoun being used pleonastically. The *P(pattern | pleonastic)* term computes the probability of observing a particular pattern if the "*it*" in question is pleonastic. We now turn to the details of these patterns.

3.1.3.2 The patterns

The sentence patterns in which we look for pleonastic ITs are of three kinds which we call the **adjective pattern, the passive pattern**, and the **S pattern**:

- *adjective pattern*: It (BE form) adjective ...
- *passive pattern*: It (BE form) passive S/SBAR ...
- *S pattern*: It ... S/SBAR

See Webb (1979) for many other kinds of anaphora.

The "*BE form*" includes any tense of the verb BE (realized as *is*, *are*, and *am*), and any combinations of modal operators and BE (e.g. *could be*, *might be*, *would be*, and so on). The "*S/SBAR*" means we expect a subordinate clause to follow.

The *adjective pattern* looks for sentences like (3.1):

3.1 ***It*** is important for the two companies to meet.

The *passive pattern* would match sentences like (3.2):

3.2 ***It*** is said that the two companies would merge.

The *S pattern* is the most general pattern and simply checks to see if there is a subordinate clause following the "*it*" in question. The pleonastic use of "*it*" in this pattern is illustrated by sentence (3.3).

3.3 ***It*** is a shame their meeting never took place.

The pattern probability in equation (3.40) now translates into *P(adjective pattern | pleonastic it)*, *P(passive pattern | pleonastic it)*, and *P(S pattern | pleonastic it)*.

3.2 Inside the equations

In this section, I will examine each component of equations (3.19) and (3.34) presented in the previous section (components appearing in both equations are explained once). I have outlined the intuitive meanings of these components in that section. Here we will look at them in a little more detail and provide examples to make them more concrete.

3.2.1 Gender/Number/Animacy information – $P(\rho \mid W_a)$

In this probability ρ is the pronoun and W_a is the word in the antecedent. Unlike the previous algorithms we examine in Chapter 2 where this information is somehow available to the resolution system by human judgement or by a lexicon, we obtain the information through this probability. This probability answers the question "What is the probability of using this pronoun ρ given that W_a is in the antecedent?" For a word to be the correct antecedent, it needs to agree

with ρ in gender and number. Thus we expect this probability to be high when the members of the pair (ρ, W_a) agree in their gender/number/animacy feature and low otherwise. Table 3.3 shows some raw statistics related to the gender feature.

ρ	W_a	$P(\rho \mid W_a)$
HE	Mr.	0.8828
	James	1
	president	0.8
	company	0
	Mary	0
SHE	Mrs.	0.9136
	Mary	1
	spokeswoman	1
	company	0
	Mr.	0
IT	company	0.9070
	spokesman	0
	stock	1
	president	0.2
	Mary	0
	team	0.5
	market	1
THEY	judges	1
	company	0.0465
	managers	1
	team	0.5
	market	0

Table 3.3 Gender/Number/Animacy – $P(\rho \mid W_a)$

This is collected from our small training data (47415 words). This means the frequency counts for some words can be low. For example, in Table 3.3, $P(HE \mid James) = 1$ is because there is only one occurrence of the word "James" and it is referred to by a HE. In a large corpus, one

usually does not expect to see such "*perfect*" probabilities as 1 and 0. How the counts are collected is explained in the implementation section §3.3.

One interesting figure in Table 3.3 is the one for the word "*team*". This is one those words that has a "*collectiveness*" property, i.e. they can be referred to either by a singular pronoun like "*it*" or by a plural one like "*they*". I will return to this issue in the implementation section §3.3.

3.2.2 Syntactic prominence – $P(G_w \mid G_p)$

This probability encodes the relationship between the grammatical role of the antecedent and that of the pronoun. We have seen various uses of this factor in the algorithms we presented in the previous chapter. Here, it is used probabilistically. It is used to capture the parallelism between the antecedent and the anaphor. The special property of being a subject (i.e. the most salient role) is also captured. One expects pronouns in subject positions to favor subject antecedents more than object antecedents. One also expects subjects to prefer subjects more than objects prefer subjects.

In our experiments, we recognize seven grammatical roles among which are *unmarked subject* (**UMSBJ**), *embedded subject* (**ESBJ**), *noun phrase subject* (**NPSBJ**), and *object* (**OBJ**). The tables below show $P(G_w \mid G_p)$ for these four grammatical roles. The whole 7x7 table is shown in the appendix.

Unmarked subjects (UMSBJ) are subjects of sentences, *embedded subjects* (ESBJ) are subjects of clauses, and *noun phrase subjects* (NPSBJ) are noun phrases embedded in another noun phrase which is a subject. In the following sentence (3.4):

3.4 *Joe's father* who *Mary* adores is a wonderful man.

"*Joe's father*" is an unmarked subject, "*Mary*" is an embedded subject, and "*Joe*" is a noun phrase subject. The coreference probabilities among these four grammatical roles are show in Table 3.4. These statistics are also from the training data.

$G_w \setminus G_p$	UMSBJ	ESBJ	NPSBJ	OBJ
UMSBJ	0.6714	0.4414	0.8	0.1967
ESBJ	0.1362	0.2920	0.1	0.2295
NPSBJ	0.1221	0.1149	0.1	0.1639
OBJ	0.0423	0.0759	0	0.2951

Table 3.4 $P(G_w \mid G_p)$ – four categories

The probabilities in Table 3.4 do agree with the intuitions. One thing that is worth pointing out is the function of noun phrase subjects (NPSBJ) such as the "*Joe*" in sentence (3.4). Compare sentence (3.4) with sentence (3.5).

3.5 The father *of Joe* who Mary adores is a wonderful man.[1]

These two sentences essentially mean the same thing. But the information values of "*Joe*" in the two sentences are different. The "*Joe*" in (3.4) seems more salient than the "*Joe*" in (3.5) which is inside a prepositional phrase. But the "*Joe*" in (3.4) is not a subject per se. One may wonder whether this fine distinction of NPSBJ from UMSBJ/ESBJ could be *statistically* significant. We did experiments in which we let noun phrases *inherit* the "*subjecthood*" of their parents if their parents are subjects. In other words we only distinguish between *unmarked subjects* and *embedded subjects*. The resulting $P(G_w \mid G_p)$ are shown in Table 3.5.

$G_w \setminus G_p$	UMSBJ	ESBJ	OBJ
UMSBJ	0.7718	0.5111	0.3115
ESBJ	0.1602	0.3407	0.2787
OBJ	0.0437	0.0730	0.2951

Table 3.5 $P(G_w \mid G_p)$ – subjecthood inheritance

There is a clear contrast between the two tables. Experimental results show that distinguishing NPSBJ is better than collapsing it with UMSBJ/ESBJ.

---

[1] The potential ambiguity involving who *Mary* adores (*Joe* or *his father*) is not the issue here and does not affect the interpretation of "*Joe*".

3.2.3 Discourse salience – $P(a \mid M_a, S_\rho)$ and $\frac{P(M_a \mid a, S_\rho G_\rho)}{P(M_a)}$

Recall that

- M_a is the mention counts of an antecedent
- S_ρ is the sentence number of the sentence in which ρ occurs. Sentences in a discourse segment are numbered sequentially starting from 1.
- G_ρ is the grammatical role of ρ

Both of these two terms approximate discourse topics but in different forms. $P(a \mid M_a, S_\rho)$ is used in the basic model and $\frac{P(M_a \mid a, S_\rho G_\rho)}{P(M_a)}$ is used in the syntactic-prominence model.. The idea is that noun phrases that are mentioned repeatedly are likely to be the topic and thus have more discourse salience which in turn makes them more likely to be pronominalized. We also need to take into consideration the position in the discourse where we find the pronoun. The nearer the end the discourse segment a pronoun occurs, the more probable it is that its antecedent has been mentioned several times.

To avoid the sparse data problem, the mention counts and the sentence numbers are *bucketed*. A portion of $P(a \mid M_a, S_\rho)$ used in the basic model is shown in Table 3.6.

M-bucket \ S-bucket	S=1 (1)	S=2 (2 – 3)	S=3 (4 – 7)	S=4 (8 – 12)	S=5 (13 – 20)
M=1 (1)	0.3106	0.0594	0.0275	0.0242	0.0282
M=3 (3 – 4)	0.4	0.5238	0.3494	0.2323	0.2522
M=6 (12 – 16)	0	0	0.1579	0.2174	0.3478
M=8 (23 – 29)	0	0	0.2174	0.3333	0.6

Table 3.6 Mention counts $P(a \mid M_a, S_\rho)$ – used in the Basic model

In Table (3.6), *M-bucket* means bucket for mention counts and *S-bucket* means bucket for the sentence. In the first row, the first number is the sentence bucket number and the actual range is in the parentheses following it. In the first column, the first number represents the mention counts bucket and the actual range is shown in the parentheses. Pronouns early in the discourse pick antecedents with lower mention counts since there have not been many entities introduced yet. This can be seen by the decrease in rows 2 (M=1) and 3 (M=3). Here in the second sentence (S=2), noun phrases mentioned 3 – 4 times are already favored over the ones that are mentioned only once (0.5238 > 0.4). As the discourse develops, more entities are introduced and those with high mention counts, being more salient, are good candidates for pronominalization. This can be verified in rows 4 (M=6) and 5 (M=8).

Table 3.7 shows some of the numbers computed by $\frac{P(M_a \mid a, S_p, G_p)}{P(M_a)}$. Since this is a ratio, the numbers do not reflect a probability distribution. The same conventions as those in Table 3.6 are used.

		M = 1 (1)	M = 2 (2)	M = 3 (3 – 4)	M = 5 (8 – 11)
S = 3 (5 – 7)	UMSBJ	0.6372	4.7076	4.1747	5.5546
	OBJ	1.2743	4.2369	4.6965	6.2068
S = 5 (13 – 20)	UMSBJ	0.5947	2.8246	1.8786	4.347
	OBJ	0.4551	3.6316	2.6837	5.3563
S = 6 (21 – 29)	UMSBJ	0.5461	2.5048	1.6664	6.4366
	OBJ	0.4673	4.2369	2.6837	9.5891

Table 3.7 Mention counts $\frac{P(M_a \mid a, S_p, G_p)}{P(M_a)}$ — used in the Syntactic-prominence model

Regardless of the grammatical role of the pronoun, a similar trend observed in Table 3.6 is also present here (i.e. high mention counts are preferred toward the end and low mention counts are

more possible at the beginning). In this table we also see the difference between subjects and objects. Subjects, being already more salient than objects, do not need as salient (in terms of *topichood* or high mention counts) a referent as do objects. We see most clearly in the last column (M = 5), the numbers for objects are higher than the numbers for unmarked subjects; whereas in the column (M = 1) the converse is observed.

3.2.4 Sentence recency — $P(S_w \mid a, f_p)$

This probability reflects preference difference between *intrasentential* antecedents and *intersentential* ones. It is conditioned on the form of the pronoun f_p. Given that reflexive and possessive pronouns tend to be "locally" bound, we expect them to favor intrasentential antecedents more than regular pronouns do. Generally, in term of discourse salience, the entities linearly closer to the pronoun (e.g. those in the same sentence) are more salient than those in previous sentences since they are more "accessible". Table 3.8 confirms these intuitions. In the table $S_w = 0$ means the antecedent is in the same sentence as the pronoun (i.e. intrasentential), $S_w = 1$ means the antecedent occurs in the immediate preceding sentence from the pronoun, and so on counting backwards.

f_p	S_w	$P(S_w \mid a, f_p)$
Reflexive	0	1
	1	0
	2	0
Possessive	0	0.9115
	1	0.0643
	2	0.0188
Regular	0	0.6003
	1	0.3634
	2	0.0296

Table 3.8 Sentence recency – $P(S_w \mid a, f_p)$

3.2.5 Syntactic constraints – $P(d_a \mid a, f_\rho)$

This probability measures the distance between the antecedent and the pronoun. But this is not some arbitrary distance measure. The distances are obtained by running Hobbs' algorithm on the input data. As I have discussed in Chapter 2, Hobbs' algorithm observes the binding principles developed in syntactic theories of anaphora. Hobbs' algorithm rules out syntactically impossible candidates. By using this algorithm to collect the candidates, we're, in fact, incorporating the binding constraints. Also, as Hobbs' algorithm starts from the pronoun and works "backwards" in the current sentence and then searches the previous sentences, we get a measure of proximity. The distance is conditioned on f_ρ. Recall that f_ρ is the from of ρ which is either *reflexive*, *possessive*, or *regular*. They were originally assumed to be independent, i.e. we only compute $P(d_a \mid a)$. However the statistics are contrastive enough to make this dependence. The probabilities of both are shown in Table 3.9 and Table 3.10.

d_a	$P(d_a \mid a)$
1	0.6142
2	0.1097
3	0.0867
4	0.0478
5	0.0274
⋮	⋮

Table 3.9 Hobbs' distance – $P(d_a \mid a)$

f_p	d_a	$P(d_a \mid a, f_p)$
Reflexive	1	0.7857
	2	0.1429
	3	0
	4	0.0714
Possessive	1	0.6836
	2	0.1394
	3	0.0777
	4	0.0322
Regular	1	0.5760
	2	0.1763
	3	0.0929
	4	0.0552

Table 3.10 Hobbs' distance – $P(d_a \mid a, f_p)$

In both tables the probabilities drop fast as antecedents move farther away from the pronoun. In Table (3.10) where the distance depends on f_p, it correctly predicts that the antecedent of a reflexive pronoun is closer to it than is the antecedent for a possessive pronoun which is then closer than for a regular pronoun. Also, the probabilities for regular pronouns do not drop as fast as those for reflexive pronouns. The contrast is seen more clearly in Figure 3.1 where Table 3.10 is graphed. Using $P(d_a \mid a, f_p)$ instead of $P(d_a \mid a)$ improves the overall performance.

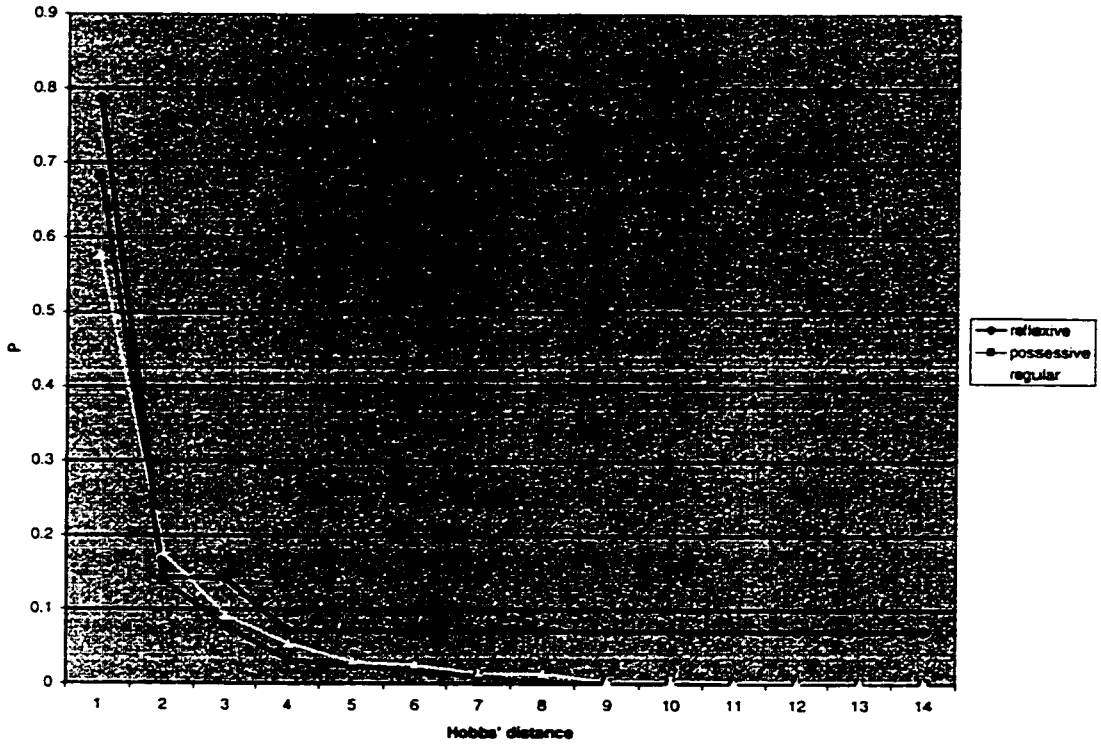


Figure 3.1 $P(d_a \mid a, f_p)$

3.2.6 Lexical semantics – $\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$

This ratio gives the likelihood of observing the antecedent word W_a under the head of the pronoun h. Recall that

- t is the type of the word W_a and is always NP
- l is the type of the head h

Usually, l is VP if the pronoun is in the object position and is S if the pronoun is in the subject position. Those two cases are depicted in Figures 3.2 and 3.3.

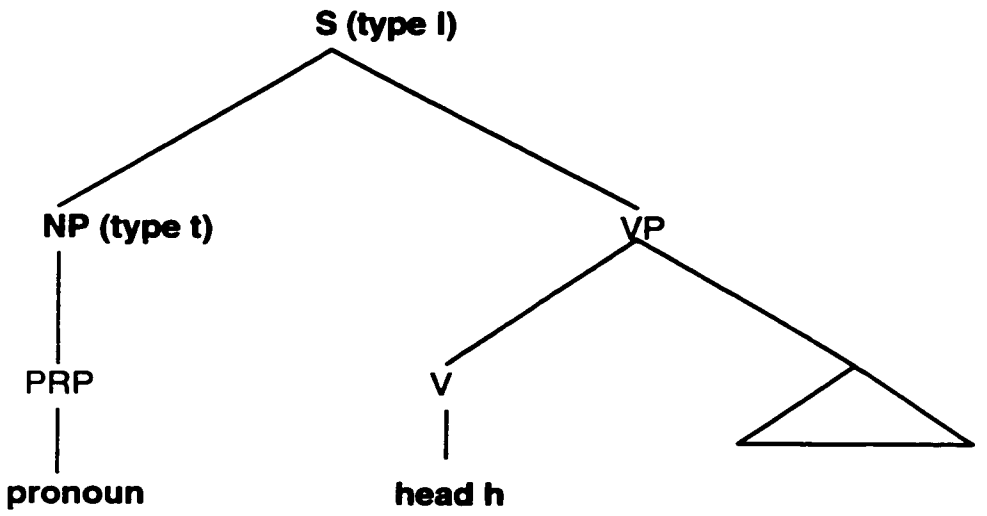


Figure 3.2 Example of S type

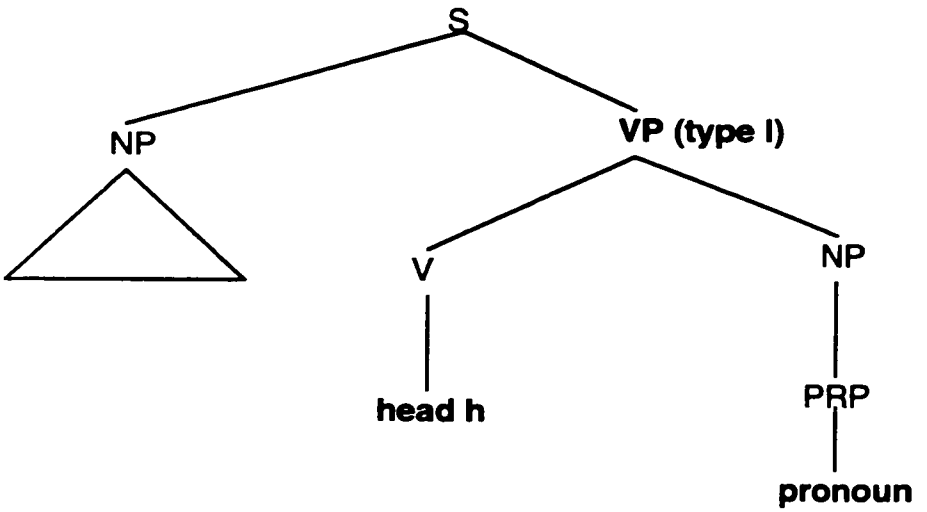


Figure 3.3 Example of VP type

Simply put, this ratio has the lexical semantics of the pair (word, head). Observe that in the l = VP case (Figure 3.3) this information represents selectional restriction. For example, for the "*it*" in (3.6):

3.6 John baked a pizza and ate ***it***.

what this term computes is *P(candidate word | eat, VP, NP)*. In this case, there are only two possible candidates, "*John*" and "*pizza*". Hence *P(John | eat, VP, NP)* and *P(pizza | eat, VP, NP)* are computed. Since the verb "*eat*" usually selects food as its direct object, one would like to see that

P(pizza | eat, VP, NP) > P(John | eat, VP, NP)

Table 3.11 shows some of the good information provided by this ratio.

type (l)	head (h)	word (W_a)	$\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$
S	said	spokesman	14.662
		agency	3.1815
		employees	0.9841
		consideration	0.1158
	explained	he	9.3326
		agency	1.1618
		game	0.7268
		trading	0.3956
VP	build	stadium	25.4494
		team	7.5540
		voters	0.8221
		idea	0.6407
		promotion	0.5852

Table 3.11 Lexical semantics — $\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$

3.2.7 Pleonastic pattern statistic – *P(pattern | pleonastic)*

As the name suggests, this statistic computes how probable a particular pattern is observed for the pleonastic us of *IT*s. Recall that we identify three patterns: *adjective*, *passive*, and *S* patterns. In actual implementation, there are two possibilities in the adjective pattern. It is observed that some adjectives are more likely to signal the use of pleonastic "*it*" than others are.

In RAP, a set of such "*typical*" adjectives is hand picked and maintained by the system. We also try to identify such adjectives, but not by hand (see §3.3.3.4 for implementation details). In essence, through an automatic program we construct a mini-dictionary containing those adjectives that are highly likely to occur with pleonastic *IT*s. The list of all adjectives thus learned can be found in the appendix. The statistics P(pattern | pleonastic) is shown in Table 3.12.

pattern	**P(pattern | pleonastic)**
adjective in dictionary	0.5
adjective not in dictionary	0.1765
passive pattern	0.0294
S pattern	0.1765
None of the above	0.1176

Table 3.12 Pleonastic pattern – *P(pattern | pleonastic)*

3.3 Implementing the algorithm

In this section, I will discuss the details of the implementation. I will first describe how Hobbs' algorithm is modified and implemented in the system (section §3.3.1). How each statistic is computed is presented in section §3.3.2. One of the statistics, that of the gender/number/animacy information can be improved using the techniques described in section §3.3.3. Finally, section §3.3.4 shows how these statistics are used to resolve pronouns in test data.

3.3.1 Implementing Hobbs' algorithm

There are a few assumptions Hobbs algorithm makes about syntax. Most notably is that the algorithm depends on the existence of an $\overline{\text{N}}$ parse tree node which is absent from the Penn Treebank parse trees[1] (there are other differences between Hobbs trees and Penn trees, but this is

[1] We use as our training and test data the Penn Treebank WSJ corpus.

the most important one as far as the problem at hand is concerned). We have implemented a slightly modified version of Hobbs' algorithm for the Treebank parse trees. In addition, we transform our trees under certain conditions to meet Hobbs' assumptions as much as possible. These modifications are discussed below.

The original Hobbs algorithm (Hobbs 1976) did not deal with reflexive pronouns. The governing domain it selects is the first *NP* or *S* node going up the parse tree from the NP node immediately dominating the pronoun. For reflexive pronouns, which are bound within their governing domains (see §1.1 for binding principles), this choice of governing domain does not always work. Consider the parse tree in Figure 3.4. NP_1 is the first NP node from NP_0 up the tree and there is no possible antecedent for the reflexive "himself" in its domain (i.e. the subtree dominated by NP_1). The governing domain is the S node up from NP_1. For this reason, we pick the first *S-type* node (S, SBAR, etc.) going up the tree from the pronoun as the minimal governing domain. If the pronoun is reflexive, only the subtree within the minimal domain is searched. The searches for an antecedent of a pronoun (reflexive or not) are done in Hobbs' fashion.

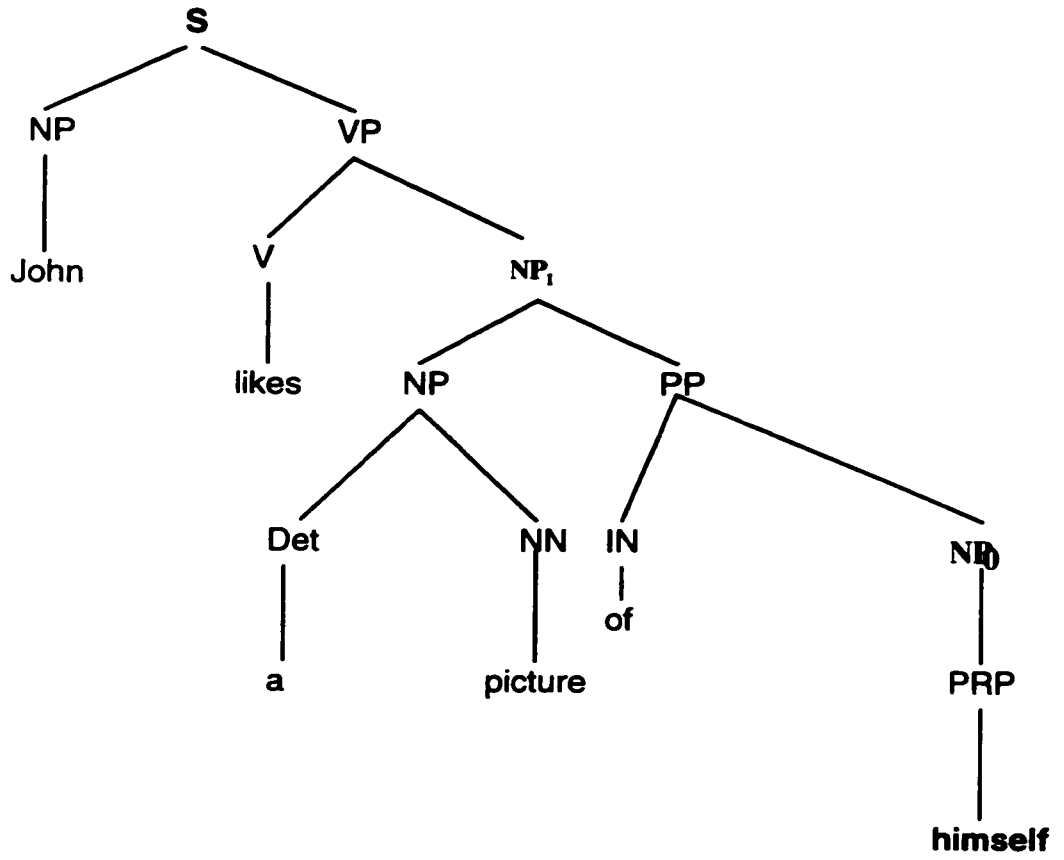


Figure 3.4 Minimal domain

The NP structure in the Penn Treebank are somewhat simplified and the trees usually have more NPs than their corresponding representations using Hobbs' trees. In particular, there are cases where an NP is immediately dominated by another NP. In general, removing the parent NP does not alter the syntactic structure and makes the tree more tuned toward Hobbs' tree. When we spot (sub)trees in form (3.5A) we transform it into (3.5B) as shown in Figure 3.5.

In our parse trees, the infinitive TO clauses usually appear under an S node. That S node does not really define a minimal domain for the pronouns under it and we can remove it to help the program find the correct domain. In general, trees like (3.6A) are transformed into (3.6B)

For each pronoun we run this modified Hobbs' algorithm repeatedly until it has proposed n (=25 in our experiments) candidates. The i^{th} candidate is regarded as occurring at *Hobbs' distance* $d_H = i$..

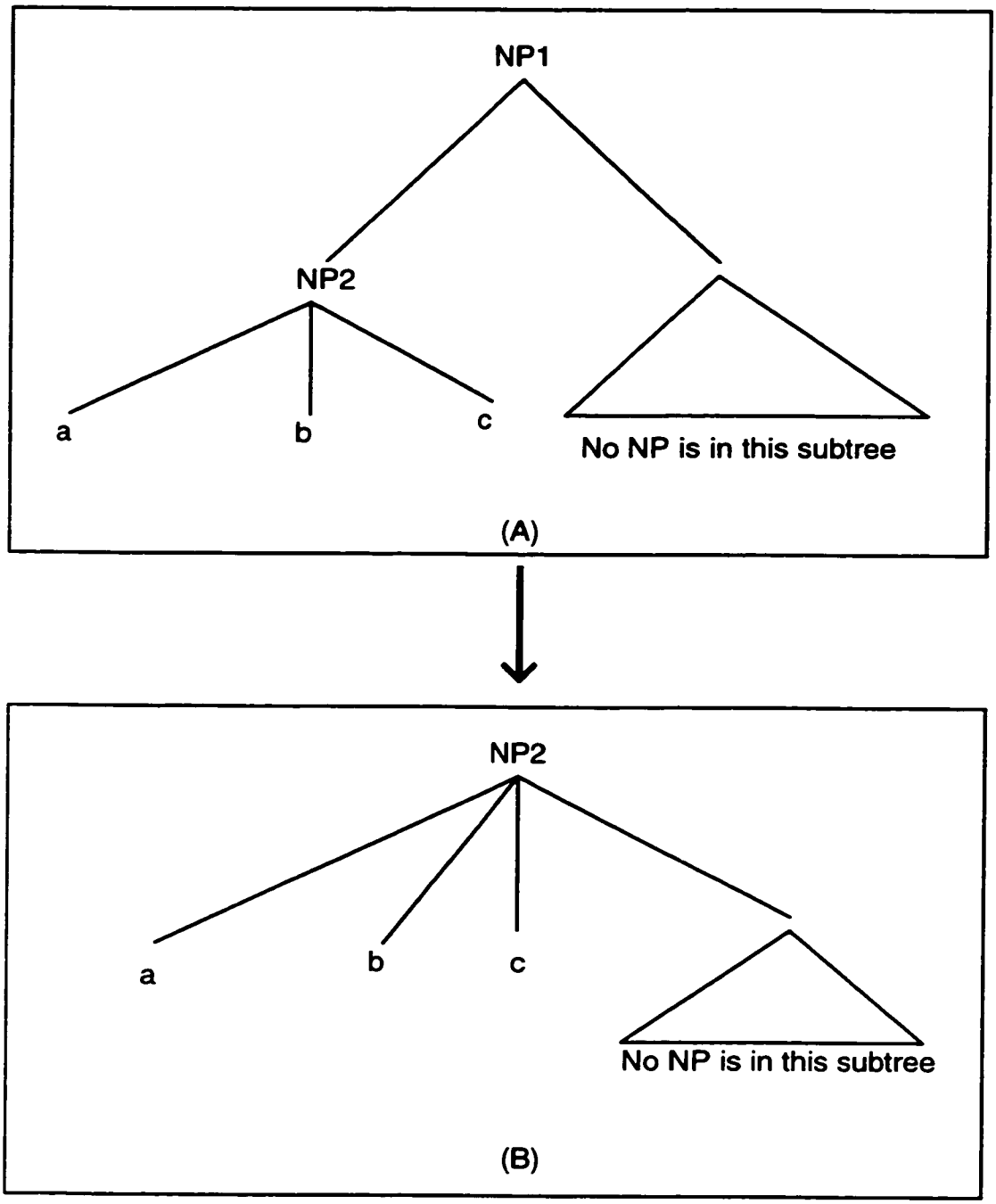


Figure 3.5 Collapsing a parse tree

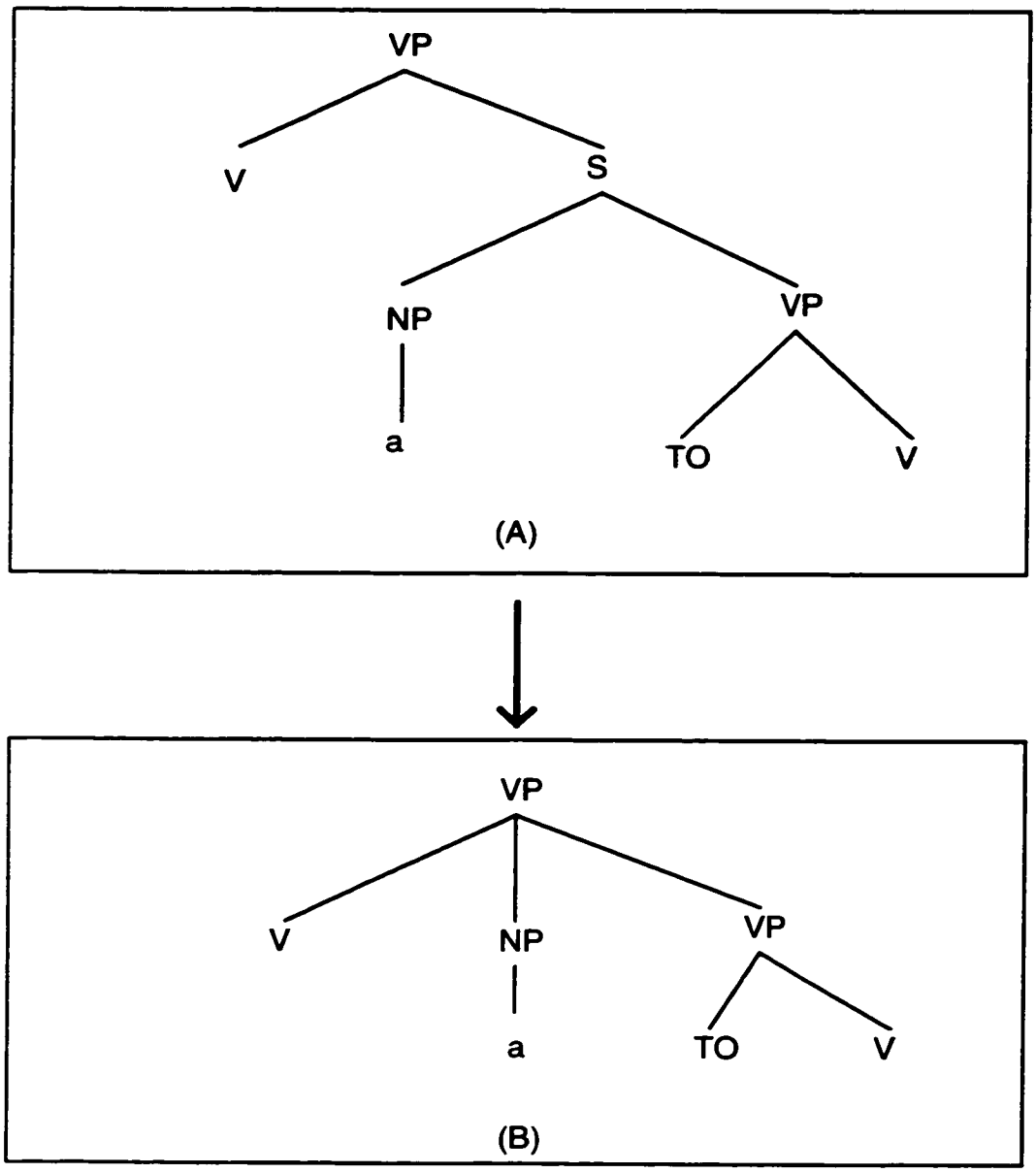


Figure 3.6 Raising a parse tree

3.3.2 Collecting statistics

We use a small portion of the Penn Wall Street Journal (WSJ) Treebank as our training corpus. The corpus is manually marked with coreference indices and referents' mention counts (see section §3.4.1 for details on corpus annotation).

After running Hobbs' algorithm on the training data, we will have gathered all information necessary to compute the statistics. Let *C* denote *Count*. The distance measure is computed as follows:

$$P(d_H = i \mid a, f_\rho) = \frac{C(\text{antecedent found at Hobbs' distance i for pronoun of form } f_\rho)}{C(\text{total number of correct antecedents for pronoun of form } f_\rho)}$$

$$= \frac{C(\text{antecedent found at Hobbs' distance i for pronoun of form } f_\rho)}{C(\text{total number of anaphoric pronouns of form } f_\rho)}$$

After we have identified the correct antecedents, it is a simple counting procedure to compute $P(\rho \mid a, W_a)$ where W_a is the correct antecedent for the pronoun ρ. The pronouns are grouped by their gender and number. There are seven of them: *HE*, *SHE*, *IT*, *THEY*, *WE*, *I*, and *YOU*.

$$P(\rho \mid a, W_a) = \frac{C(\text{number of times } W_a \text{ occurs in the antecedent for } \rho)}{C(\text{number of times } W_a \text{ occurs})}$$

The referents range from being mentioned only once to 120 times in the training examples. Instead of computing the probability for each one of them, we group them into "*buckets*" so that M_a is the bucket for the number of times that the antecedent ***a*** is mentioned. For example, bucket 1 and bucket 2 contain those antecedents that are mentioned once and twice respectively, M_3 contains those that are mentioned three or four times, etc. with bucket size increasing with bucket number. Same bucketing scheme is applied to sentence number S_ρ. The method to compute the mention counts probabilities are:

$$P(a \mid M_a, S_\rho) = \frac{C(\text{correct antecedent mentioned } M_a \text{ times for pronoun in sentence } S_\rho)}{C(\text{proposed antecedents mentioned } M_a \text{ times for pronoun in sentence } S_\rho)}$$

$$P(M_a \mid a, S_\rho, G_\rho) =$$

$$\frac{C(\text{correct antecedent mentioned } M_a \text{ times for pronoun occupying } G_\rho \text{ position in sentence } S_\rho)}{C(\text{number of pronouns occupying } G_\rho \text{ position in sentence } S_\rho)}$$

After the correct antecedent is found for a pronoun ρ, its grammatical role is determined. We distinguish the following seven grammatical roles:

- *unmarked subject* (UMSBJ): subject of a sentence

- *embedded subject* (ESBJ): subject of an embedded clause
- *noun phrase subject* (NPSBJ): noun phrases whose parent is also a noun phrase and the parent is in a subject position
- *object* (OBJ): noun phrases following a verb
- *prepositional phrase* (PP): noun phrases embedded in a prepositional phrase that is not preposed
- *preposed prepositional phrase* (PPS): same as PP except that the prepositional phrase is preposed
- *other* (OTHER): none of the above

There is a little more that needs to be said about how UMSBJ, ESBJ, and NPSBJ are classified. The most straightforward case is when an NP in the subject position is directly dominated by the top S node. The NP in Figure 3.7 is thus an UMSBJ:

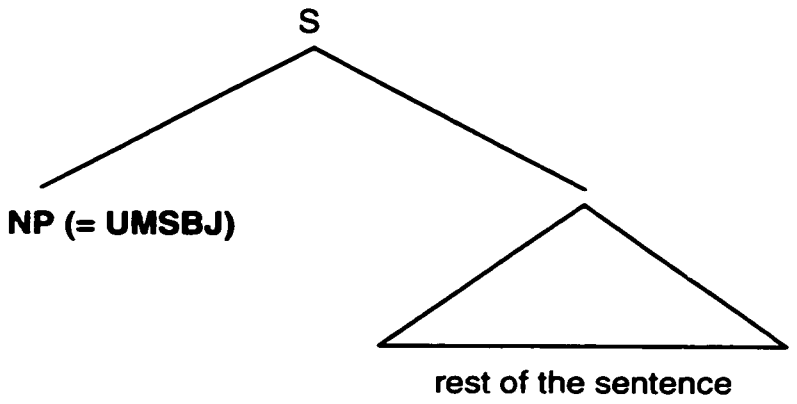


Figure 3.7 UMSBJ: case 1

An NP is also classified as UMSBJ when it is not directly under the top S node but the rest of the sentence does not contain a subject NP as in Figure 3.8A and an example sentence is provided by Figure 3.8B where '*John*" is treated as an unmarked subject. In the figure, $[S]^+$ means one or more occurrences of S-type node. When there is a subject NP in the rest of the sentence, then the

NP in Figure 3.8A is an embedded subject as shown in Figure 3.9. Figure 3.10 shows the configuration for NPSBJ. Examples of PP and PPS are shown in Figure 3.11 and 3.12 respectively. Now we are ready to compute the syntactic prominence probability:

$P(G_{Wa} \mid a, G_\rho) =$

$$\frac{\text{C(number of antecedents occupying position } G_{Wa} \text{ for pronoun occupying position } G_\rho)}{\text{C(number of pronouns taking grammatical role } G_\rho)}$$

The remaining two statistics are sentence recency $P(S_{Wa} \mid a, f_\rho)$ and pleonastic pattern statistics $P(pattern \mid pleonastic)$. They are computed as follows.

$P(S_{Wa} \mid a, f_\rho) =$

$$\frac{\text{C(number of antecedents occurring in } S_{Wa} \text{ sentence relative to } \rho \text{ for } \rho \text{ of the form } f_\rho)}{\text{C(number of pronouns of the form } f_\rho)}$$

Finally,

$$P(\text{pattern} \mid \text{pleonastic}) = \frac{\text{C(number of sentences with a pleonastic } it \text{ that match the pattern)}}{\text{C(number of pleonastic } IT\text{s)}}$$

In building a statistical parser for the Penn Treebank, various statistics have been collected (Charniak 1997), two of which are $P(w \mid h, t, l)$ and $P(w \mid t)$. To avoid the sparse data problem, the heads h are clustered according to how they behave in $P(w \mid h, t, l)$. The probability of w is then computed on the basis of h's cluster $c(h)$. Obtaining these two probabilities, we can then compute the ratio $\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$.

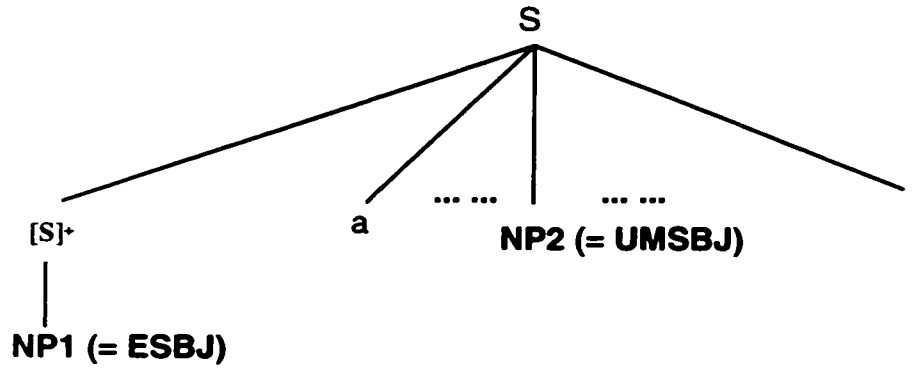
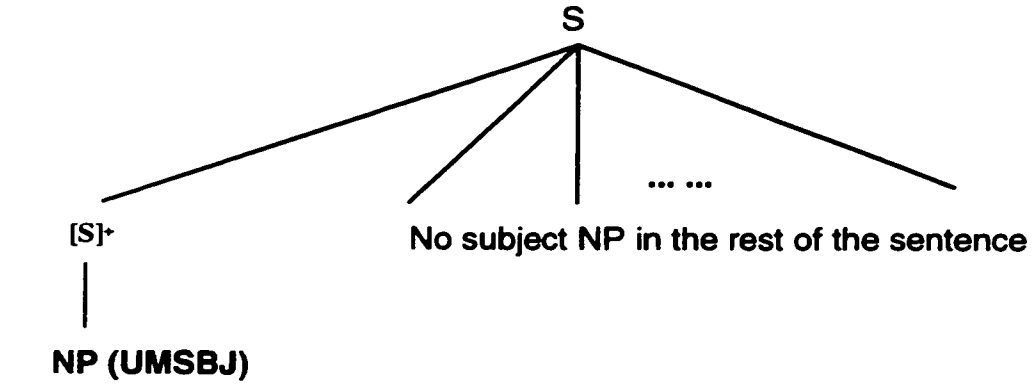
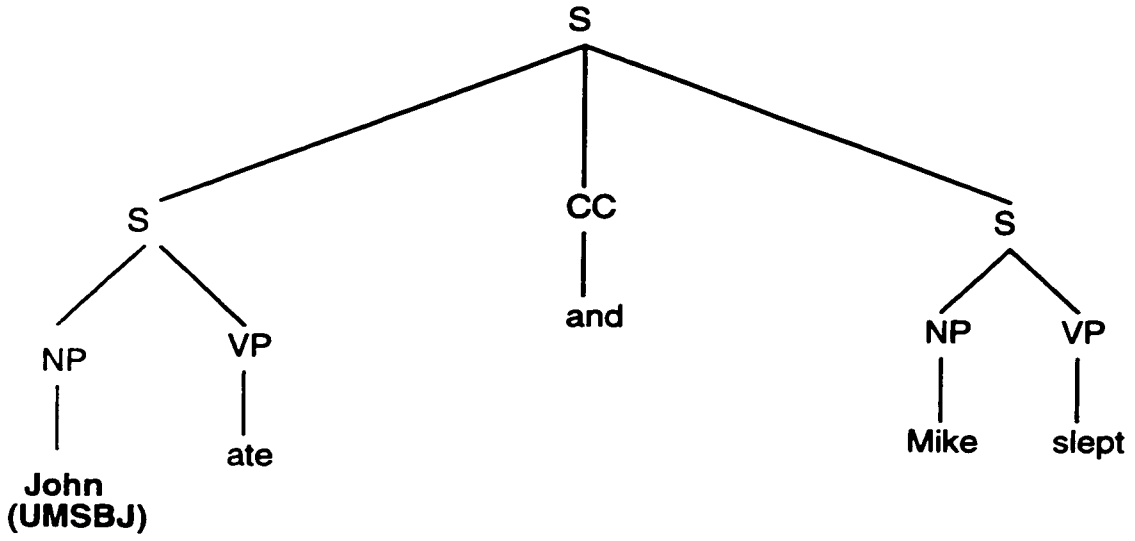


Figure 3.9 ESBJ configuration



(A) UMSBJ case 2: configuration



(B) UMSBJ case 2: example

Figure 3.8 UMSBJ: case 2

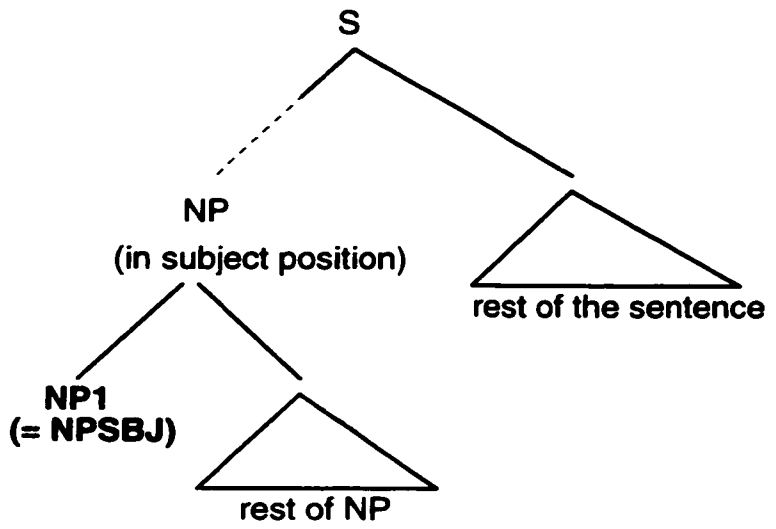


Figure 3.10 NPSBJ configuration

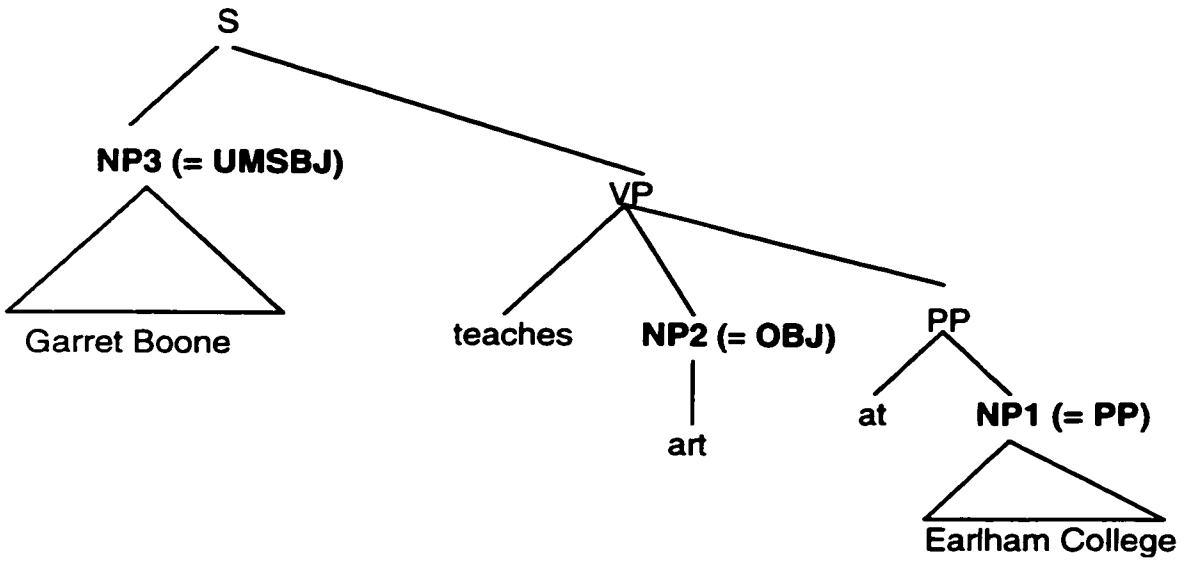


Figure 3.11 PP example

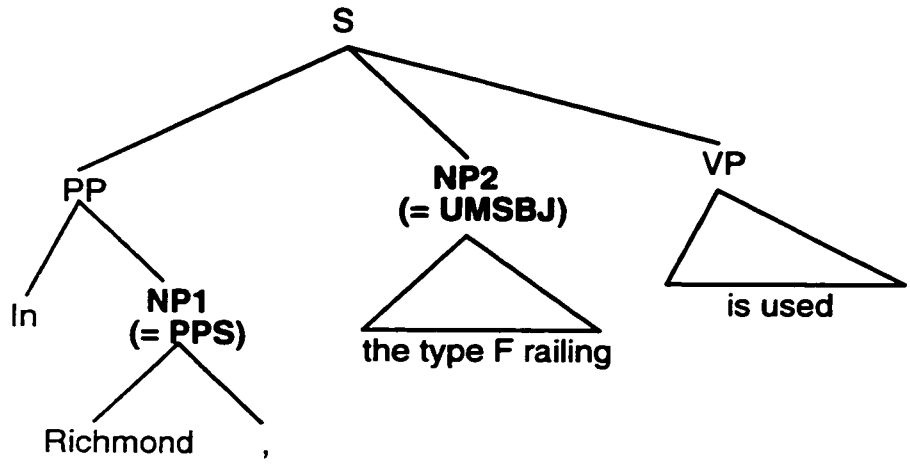


Figure 3.12 PPS example

3.3.3 Gathering more information

The statistics described in the previous section (with the exception of the word semantics $\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$) are collected from a small coreference-marked corpus (47,415 words and 1968 sentences). The gender/animacy information learned is reasonably reliable since the correct anaphoric links are established before the counts are gathered. Good as this may be, the statistic has limited coverage of words simply because the data set is limited[1]. This section presents three methods that can help overcome this problem and also give us more accurate knowledge about this feature. In section §3.2.7, I mentioned that a small dictionary of adjectives for pleonastic *IT*s are built and here I will present the way to do it automatically.

3.3.3.1 Unsupervised learning of gender information

To learn more about the gender feature, we consider an automatic method for estimating the probability that nouns occurring in a large corpus of English text denote inanimate, masculine

---

[1] We could, of course, mark up more data with pronouns correctly resolved. Needless to say, this requires a lot of manual work and isn't very appealing.

or feminine things. The method is a very simple mechanism for harvesting the kind of gender information present in discourse fragments like "***Mary slept. She slept for a long time.***" After processing the second sentence and identifying "*She*" with "*Mary*", the gender of "*Mary*" is known. To find referent/pronoun pairs, we took the 21-million-word 1987 Wall Street Journal corpus and run Hobbs algorithm on it[1]. This is a very naïve and inexpensive approach. It simply takes the first NP proposed by Hobbs' algorithm as the antecedent for the pronoun in question. The probability $P(\rho \mid w)$ is computed as before (section §3.3.2). Some statistics gathered from this learning method are in Table 3.13.

Word	P(HE)	P(SHE)	P(IT)
Company	0.0764	0.0060	0.9174
President	0.8206	0.0139	0.1654
Mr. Reagan	0.8820	0.0037	0.1142
Government	0.1172	0.0122	0.8704
Mrs. Thatcher	0.0735	0.8235	0.1029
Judge Bork	0.8820	0	0.1179

Table 3.13 Hale's good statistics

Obviously, this syntax-only pronoun resolution strategy will be wrong some of the time. As we noted in section §3.2.5, accuracy for noun phrases found at Hobbs' distance 1 is 61.42%. This shows up in the noises in the statistics it produces. Some of the *not-so-good* ones are listed in Table 3.14.

Word	P(HE)	P(SHE)	P(IT)
Spokesman	0.6075	0.0045	0.3879
Years	0.5298	0.0815	0.3886
Daughter	0.2340	0.7021	0.0638
Judge	0.7154	0.0836	0.2008

Table 3.14 Hale's noisy statistics

[1] Mr. John Hale, then an undergraduate at Brown, did this experiment and it is reported in Ge, Hale, and Charniak (1998).

Noisy as this data may be, it learns a lot more words and is still better than nothing. What we want to do next is try to have the better of both worlds. In other words, we would like a larger coverage of words than the one collected from the limited training set and at the same time, have more accurate probabilities than those computed by this method.

3.3.3.2 Near "perfect" information

We can get some very good gender information regarding inanimate objects and human objects. English has relative clause constructions. A ***relative clause*** is a sentence-like construction that typically begins with a *relative pronoun* (*who*, *whom*, *which*, etc.) and which is appended to a noun. Observe that the choice of "*who*" and "*which*" is not arbitrary. Which of the two is to be used depends on whether the noun to which the relative clause is appended is or is not human.

Having observed this phenomenon, we ten simply go through a large corpus (1-million-word) looking for occurrences of "*which*" and "*who*" (also its variant "*whom*"[1]). The nouns preceding "*which*" are classified into ***WHICH*** class meaning that they are inanimate and therefore cannot be referred to by pronouns in classes *HE/SHE/I*. Similarly noun preceding "*who*"/"*whom*" are classified into ***WHO*** class and they cannot be the antecedents for pronouns in the *IT* class.

We need to be a little careful in collecting those nouns. We check if the noun phrase preceding *which/who* begins with a *possessive construction* (either a possessive pronoun or a ('s) construction). If this is the case, we only collect nouns after the possessives. This means in the phrase "*Mr. Smith's* ***company which*** ...", only the word "*company*" is collected. In the case where there is an "of" construction under a noun phrase, only the nouns preceding the "*of*" are

---

[1] "whose" is not used because it can be attached to an inanimate object as in:
Chez. Panisse Corp. ***whose*** founder is the inventor of California cuisine cooking style hasn't subjected diners to vanilla ice cream.

gathered. Hence, in phrases like "*the **director** of XY Corp. **who** ...*", only the word "*director*" is added to the WHO class.

This simple method of constructing a WHICH class and a WHO class gives very accurate animacy information. The data does not have to have pronouns resolved and thus allows us to do this on a relatively large data set. Some examples of the WHICH class the WHO class are shown in Table 3.15.

WHICH Class	**WHO Class**
Administration	Allan
Building	Buyers
California	Wife
WHICH Class	**WHO Class**
Institute	Harry
Plan	Porter
Reform	Representative
Service	Steve
Toshiba	Traders
University	Viewers

Table 3.15 WHICH and WHO

Except for a few overlaps the lists are very clean. Most of the overlapping cases involve proper names that can be either company names or personal names. Some examples are shown in Table 3.16.

Anderson
Philip
Family
Group
Candy

Table 3.16 WHICH and WHO overlaps

There are total 71 overlapping words compared to 2111 words in the WHICH class and 1227 words in the WHO class. Given this contrast, we do not believe that putting a probability distribution on the two classes will make statistically significant contributions to the system's performance. Therefore, we choose to exclude the overlaps from both classes.

3.3.3.3 Simple transductive learning technique

The overall structure of the system has two components in it: a ***collector***, responsible for gathering various information from the training data and computing necessary statistics, and a ***resolver*** which runs on the test data, uses the collector's statistics, and resolves the pronouns in the data. A very simple transductive learning technique can be used in the resolver. After the resolution program finds an antecedent for a pronoun, it collects the antecedent/pronoun pair and *recomputes* the statistic $P(p \mid a, W_a)$ before going on to the next pronoun. In effect, the system is learning from the answer it generates. Note that in the training phase, the *collector* learns from the *correct* antecedents, whereas here the *answer* from which the system learns may not be correct. Consequently, this newly gathered information will not be as good but is still useful and our experiments show that it does improve the accuracy of resolution.

3.3.3.4 Adjectives for pleonastic IT

In the Penn Treebank, there is a type of empty nodes called the "***expletive***" (*EXP*). This marking essentially tells whether the preceding NP is or is not extrapolated. In case of a pronoun IT, an expletive null node following it identifies it as pleonastic. An example tree with an expletive null node is shown in Figure 3.13. The way to collect adjectives then becomes simple: go through the parse trees, look for ITs followed by an expletive null node, and collect the following adjective if there is any. In Figure 3.13, the adjective pattern is observed and hence

the adjective "*unclear*" is added to the dictionary. The list of all the adjectives collected this way is shown in the appendix.

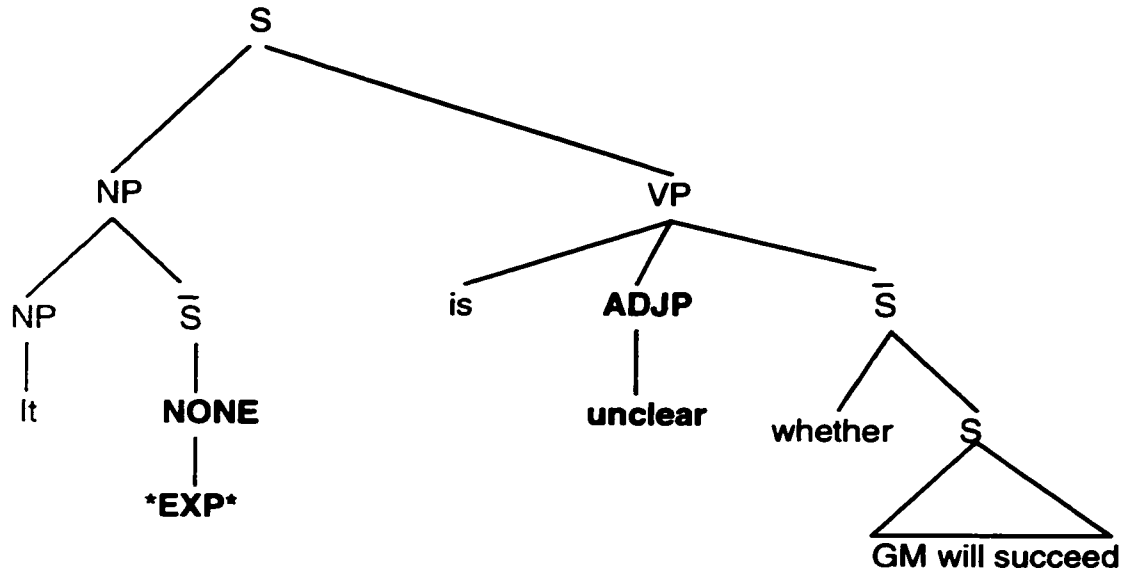


Figure 3.13 *EXP* empty nodes

3.3.3.5 Bllip99 statistics

With the statistics collected (section §3.3.2) and all of the above additional information incorporated, the resolution program was run on a 35-million-word Wall Street Journal corpus (1987-1989 provided in the Penn Treebank Project Release 2). The corpus was first parsed using a statistical parser (Charniak 2000). After that grammatical and function tags were assigned (Blaheta and Charniak 2000). Another program was then run on the data to identify full noun phrase coreferences. (Hall 2000) Empty nodes were then inserted into the parse trees. (Charniak 2000) Finally the pronoun resolver was run.

We apply a similar technique described in section §3.3.3.1 to the Bllip99 corpus, i.e. collecting *(pronoun, reference)* pairs. Bllip99 data not only gives us more words than the Hale corpus which has 21 million words, it also contains more accurate gender information because the Resolver was more equipped than the simple Hobbs' algorithm used to collect the Hale's statistics. The information collected from Bllip99 is stored and fed back into the Resolver. The idea is very similar to the Hale's statistics, only that the information is much more accurate.

The process can potentially be repeated until the gender statistic does not improve performance any more. It is very similar to the convergence condition in the Expectation Maximization (EM) learning technique. But since the Bllip99 corpus is very large (35 million)

compared to our test data (46,516 words), we don't expect the gain from repeating the process to be statistically significant and therefore this learning process is applied only once.

3.3.4 Resolving pronouns

The overall organization of the system is depicted in Figure 3.14. Note that both the training data and the test data are marked with coreference numbers and mention counts. The data on which the *additional helpers* run is just Penn parse trees with no additional markings. The information provided by the *additional helpers* is pre-collected and is stored, i.e. these helper programs are only run once and the results are stored in files which the *resolver* will read. Because of the need to run *cross validations*, the collector is run every time the resolver is run. In Figure 3.14, the implementation of the *collector* is discussed in section §3.3.2 and the *additional helpers* are described in section §3.3.3. The procedures in the *resolver* are rather straightforward. In the remaining of this section I will clarify a few issues in the resolver.

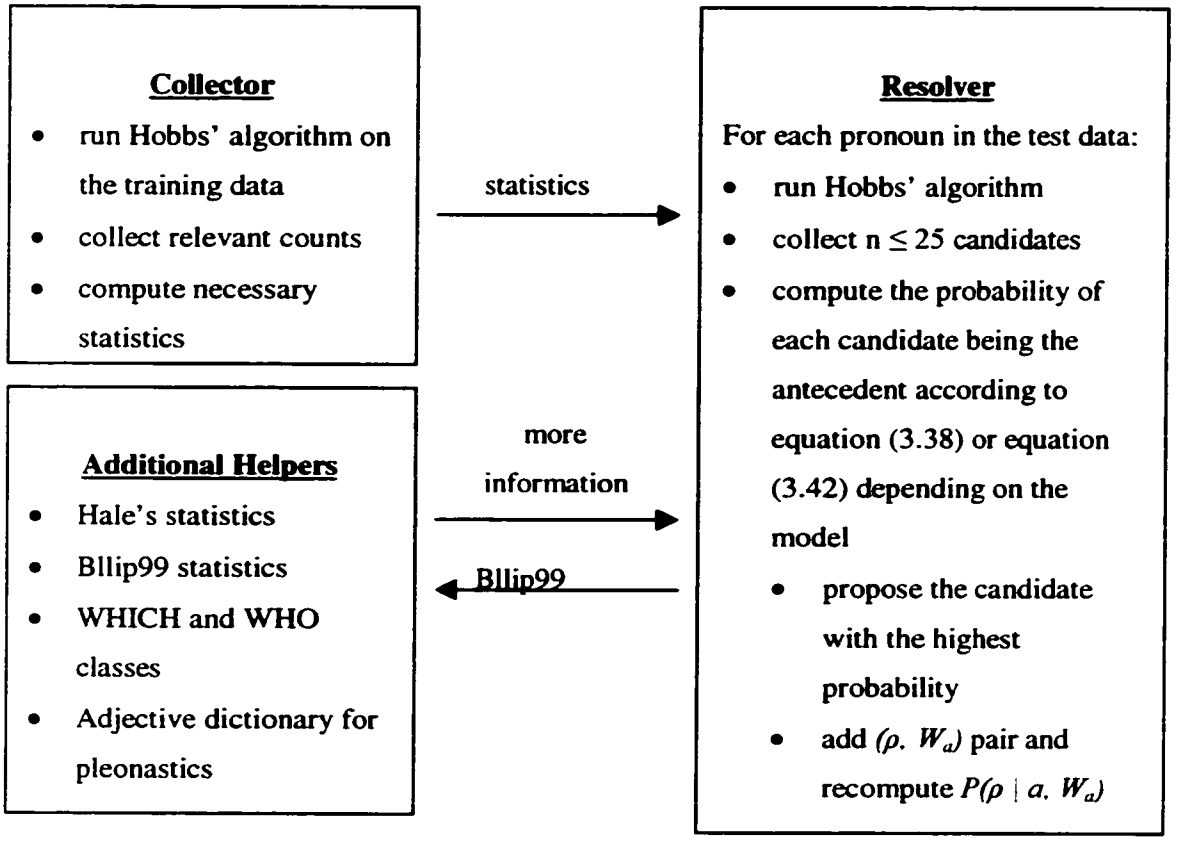


Figure 3.14 The resolution system

First of all, the proposed antecedent output by the resolver is, in most cases, a noun phrase. In case of the pronoun "*it*", if the probability of being pleonastic is the highest, then the resolver declares the "*it*" in question to be pleonastic and no antecedent is proposed.

For a particular candidate noun phrase, we need to decide which word in the phrase to use by the gender statistics $P(\rho \mid a, W_a)$. We want the word to be the most *informative* one in this noun phrase. This can be accomplished by performing the likelihood ratio test[1] (Dunning 1993).

[1] We choose Dunning's likelihood test over standard tests like Pearson's χ^2 and Z-score tests because of the small size of the training data. These common tests make the assumption of normality. This assumption

This statistic is based on the binomial or multinomial distribution and is applicable to smaller texts such as ours. The formulae to compute the likelihood ratio statistic are included in the appendix. Intuitively, the word *w* thus selected is the word most likely to be observed with a pronoun. Top 15 most likely words are shown in Table 3.17.

Word	Log-likelihood ratio
Mr.	327.31
Mrs.	321.321
Yeargin	283.53
Marie-Louise	83.6106
company	78.4301
companies	60.189
Hahn	55.3084
Dinkins	49.4246
Ms.	49.1213
Ward	43.9856
Corp.	39.2843
Inc.	38.0189
Judge	35.066
Artist	34.7795
Viacom	33.5769

Table 3.17 Top 15 most likely words

If we have never before seen any of the words in the candidate noun phrase, we simply use the prior probability of the pronoun, i.e. $P(p)$.

After *w* in the candidate noun phrase is selected, it passes through a series of simple tests depending on the pronoun in question before $P(p \mid a, W_a)$ is used.

- If the pronoun is singular (HE/SHE/IT/I), then any plural nouns are assigned a probability 0 (effectively being ruled out)

---

breaks down when comparing the rates of occurrence of rare events. The small-sized data set is composed largely of such rare events.

- **Check animacy agreement between *w* and *p* using the WHICH and WHO helper. Assign 0 probability to any candidates failing this test. In other words, candidates found in the WHO class are ruled out for pronouns in the IT class and candidates found in the WHICH class are ruled out for pronouns in the HESHE/I class.**

The plural pronouns are exempted from these tests because plural pronouns can be either animate or inanimate. See sentences in 3.7 for example.

3.7a *John and Mary* went to a party and *they* had a lousy time.
b John bought *ten books* and read *them* all.

They are also exempted from the number test because there are certain singular *collective* nouns in English which can be referred to by the plural pronouns. An example is in sentence 3.8:

3.8 *The company* said *they* would go public next month.

However, there are not many such singular collective nouns, which means given any singular noun, its chance of having the "*collectiveness*" property is not very large. Instead of manually drawing up a comprehensive list of all such nouns[1], we guess this property as follows. The *collector* has all the antecedent/pronoun pairs found in the training data. The idea is that if a singular noun is never seen to corefer with a plural pronoun by the collector, our guess is that it is not collective and is thus ruled out[2]. Some of the singular nouns that survive this test are listed in Table 3.18.

company
government
administration
team

Table 3.18 Singular collective nouns

[1] This is what Mitkov did in his approach. (Mitkov 1998)

[2] One of the reasons why this heuristic is a good approximation is that both the training and the test data re from the same domain, namely WSJ. It may not work as well if applied to texts of another genre, for example, Romance novels.

After these tests are done, the collector is called to give the statistics that are needed to compute equation (3.38) or (3.42) for each candidate. The most probable candidate is then selected as the proposed antecedent.

3.4 The experiment

This section describes the experiments we did using the two models. The section begins with a brief explanation of the data set (§3.4.1). Empirical results are then shown in section §3.4.2, followed by a comparison of the two models (section §3.4.3). This section ends with a comprehensive error analysis in §3.4.4.

3.4.1 Corpus annotation

The data (both for training and testing) is a small portion of the Penn Treebank (93,931 words and 3975 sentences). It is then marked with pronoun coreference information. Every anaphoric pronoun in the corpus is assigned a reference number. The same number is also attached to its antecedent. As I have explained, not all pronouns have an anaphoric antecedent in the text. Thus every pronoun also has its *type* marked. Currently there are five types for pronouns:

- *Explicit object referents* (OBJREF). Those are anaphoric pronouns and they carry with them a coreference index and mention counts. There are 2002 pronouns in this category.
- *Multiple discontinuous referents* (NONLOCOBJ). Pronouns whose antecedents are not explicitly mentioned in text fall into this category. For example the "*they*" in sentence (3.9) refers to "*Mr. Stronach and Manfred Gingl*", but this phrase does not appear in the text.

  3.9 The company said *Mr. Stronach* will personally direct the restructuring assisted by *Manfred Gingl*, president and chief executive. Neither ***they*** nor Ms. McAlpine could be reached for comment.

Deictic pronouns are also included in this category. There are total 300 pronouns in this category.

- *Action* (ACTION). Sometimes a pronoun may refer to an action or an event that cannot be identified by a single noun phrase, as illustrated by example (3.10):

  3.10 The next morning, with a police escort, busloads of executives and their wives raced to the Indianapolis Motor Speedway, unimpeded by traffic or red lights. The governor could not make ***it*** so the lieutenant governor welcomed the special guests.

  There are 76 ACTION pronouns.

- *Environment* (ENV). A pronoun may simply refer to some conventional unspecified referent as the "it" in sentence (3.11)

  3.11 ***It*** is eleven o'clock.

  We have 5 pronouns of this type.

- *Pleonastic* (SYNTAX). The pleonastic *IT*s are marked with this type and thus do not carry a coreference number. Thee 67 pleonastic *IT*s.

We assume that the coreference relation is transitive. If phrase *A* is marked as referring to *B* and later phrase *C* is found to be coreferential with *B*, we conclude that *C* is also coreferential with *A*. All coreferential entities share the same reference number:

3.12 *Dell Computer Corp.* said *it* cut prices on several of *its* personal computer lines by 5% to 17%.
The company said *its* price cuts include a $100 reduction on *its* system 210 computer with 512 kilobytes of memory.

The phrases "*Dell Computer Corp.*" and "*The company*" denote the same entity and therefore are marked with the same reference number.

As I alluded to in the previous sections, the system has the knowledge of frequency of mention. This is because the noun phrases in the corpus are also marked with their mention counts. In a discourse segment (i.e. a file[1]), we count the number of times an entity is mentioned

---

[1] Each file contains a different story. File boundaries are discourse boundaries.

up to that point. They can be either explicitly mentioned by a noun phrase or referred to by a pronoun. The counts are accumulated if the same entity is mentioned again later in the story. For example, in the short discourse segment (3.13):

> 3.13 *Ralston Purina Co.* (1) reported a 47% decline in *its* (2) fourth-quarter earnings. *The company* (3) earned $45.2 million compared with $84.9 million a year earlier. *Ralston* (4) said *its* (5) restructuring costs include the phase-out of a battery facility, the recent closing of a Hostess cake bakery and a reduction in staff throughout *the company* (6).

The "Ralston company" entity is mentioned six times in total, three times by a pronoun and three times by a full noun phrase.

3.4.2 Empirical results

The evaluations reported in this section are those for anaphoric and pleonastic pronouns[1]. Pleonastic pronouns are simple. They are counted as being correctly resolved if they are identified as pleonastic. For anaphoric pronouns, the most straightforward case is when the system proposes a noun phrase as its answer and that noun phrase has the same reference number as the pronoun. This is obviously counted as correct. Sometimes the system selects a pronoun as its answer. Then the reference number associated with this answer pronoun is compared with that of the pronoun in question. If they agree then the answer is correct.

We first divide the data in half, 100 files for training and another 100 files for testing. The training set consists of 47,415 words, 1968 sentences, and 1344 pronouns. The test set has 46,516 words, 2007 sentences, and 1119 pronouns. The basic model using equation (3.38) which is reproduced here as (3.44) achieves accuracy 87.8% for anaphoric HE/SHE/IT pronouns.

---

[1] Other types of pronouns are outside the scope of this study and are excluded from evaluations.

$$F(\rho) = \underset{a}{\operatorname{argmax}} P(A(\rho) = a \mid \rho, \vec{d}_H, \vec{W}, h, t, l, \vec{M}, S_\rho, f_\rho, pattern) \quad (3.43)$$

$$= \underset{a}{\operatorname{argmax}} P(d_a \mid a, f_\rho) P(\rho \mid a, W_a) \frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)} P(a \mid M_a, S_\rho) P(pattern \mid a) \quad (3.44)$$

The syntactic-prominence model which uses equation (3.42) reproduced here as (3.46) achieves accuracy 90.7%. These results are summarized in Table 3.19.

$$F(\rho) = \underset{a}{\operatorname{argmax}} P(A(\rho) = a \mid \rho, \vec{d}_H, \vec{W}, \vec{M}, S_\rho, f_\rho, G_\rho, \vec{G}_w, \vec{S}_w, ptn) \quad (3.45)$$

$$= \underset{a}{\operatorname{argmax}} P(\rho \mid a, W_a) P(d_a \mid a, f_\rho) P(G_{w_a} \mid a, G_\rho) P(S_{w_a} \mid a, f_\rho) \frac{P(M_a \mid a, G_\rho, S_\rho)}{P(M_a)} P(ptn \mid a) P(a) \quad (3.46)$$

Test data	Anaphoric HE/SHE/IT
Basic model	87.8%
Syntactic-prominence model	91.3%

Table 3.19 HE/SHE/IT results on test data

The two models perform equally on the pleonastic ITs. There are total 33 such occurrences. Both models achieve precision 100% and recall 54.55%. See Table 3.20

Test data	Pleonastic It
Precision	100%
Recall	54.6%

Table 3.20 Pleonastic *IT*s in test data

It is quite clear that about half of the pleonastic usage of ***IT*** are not recognized. This is because those sentences do not seem to fall into any particular pattern. Some examples are:

- ***It*** could take years for the new Polish government to fully use the aid effectively.
- The House passed legislation designed to make ***it*** easier to block airline leveraged buy-outs.

- Usually *it* is large investors initiating a buy or sell in Chicago.

We could, of course, design ad-hoc patterns to account just for these cases. But doing so would require a knowledge of the test data since in order to cover all the pleonastic patterns to achieve 100% recall, we would have to know what sentence patterns occur in the test data.

The overall performance of the two models on the test data is summarized in Table 3.21.

Model	a. HE/SHE /IT	b. Pleonastic IT	c. I	d. THEY/ WE	e. Anaphoric (a, c, d)	f. Overall (a,b,c,d)
Basic	87.8%	(100%,54.6%)	88.9%	71.2%	83.4%	82.3%
Syntactic-prominence	91.3%	(100%,54.6%)	94.4%	76.7%	87.2%	86.0%

Table 3.21 Performance on the test data

Notice that pronouns in the **YOU** class do not show up in the table. This is because all occurrences of YOU-pronouns in the test data are *deictic*. In the training data, one out of the 52 YOU-pronouns is anaphoric. In Table 3.21, column (a) is the accuracy for anaphoric HE/SHE/IT pronouns and column (b) is for pleonastic ITs, column (c) shows accuracy for the anaphoric I-pronouns, column (d) is for anaphoric THEY/WE-pronouns, column (e) is the performance on all anaphoric pronouns (i.e. summary of columns (a), (c), and (d)) and the last column takes into account of all pronouns, anaphoric and pleonastic.

We then run a ten-way *cross validation* where we reserve 10% of the corpus for testing and use the remaining 90% for training. The results for anaphoric HE/SHE/IT are shown in Table 3.22.

Cross-Validation	Average Success Rate
Basic model	88.1%
Syntactic-prominence model	92.2%

Table 3.22 Cross validation results on HE/SHE/IT

We see that the syntactic-prominence model performs consistently better than the basic model.

Before I compare the two models (which is the focus of next section §3.4.3), I would like to show some incremental results. These experiments are done in order to find the relative importance of each factor (i.e. individual statistics in the equations (3.44) and (3.46)) in pronoun resolution. We run the program "incrementally", each time using one more probability term in the equations. The results are shown in Tables 3.23 and 3.24.

Factor	Statistic	Average Success Rate
Syntax	$P(d_a \mid a, f_p)$	67.7&
Gender/Number/Animacy	$P(p \mid a, W_a)$	80.2%
Lexical semantics	$\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$	82.4%
Discourse topic/salience	$P(a \mid M_a, S_p)$	88.1%

Table 3.23 Incremental results – Basic model

Factor	Statistic	Average Success Rate
Syntax	$P(d_a \mid a, f_p)$	67.7&
Gender/Number/Animacy	$P(p \mid a, w_a)$	81.2%
Syntactic prominence	$P(G_w \mid a, G_p)$	86.5%
Sentence recency	$P(S_w \mid a, f_p)$	87.2%
Discourse topic	$\frac{P(M_a \mid a, S_p, G_p)}{P(M_a)}$	92.2%

Table 3.24 Incremental results – Syntactic-prominence model

It is clear from both tables that the gender/number/animacy information is a significant contributor to the system. It gives 12.5% increase to the performance. Another big factor in both models is the mention counts statistic which has average contribution of 5.6%. In the basic model, the lexical semantic statistic $\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$ adds a marginal 2.2% whereas in the syntactic-prominence model, the grammatical role and the sentence recency statistics contribute 5.5%. I will now turn to the comparison of the two models.

3.4.3 Comparing the two models

The two models are both set up in the same statistical framework. They differ in the discourse context they look for, in other words, in different anaphora resolution factors they utilize. The syntactic-prominence model uses more information and is in a way a superset of the basic model. However, notice that the syntactic prominence model does not have the statistic $\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$, i.e. it is blind to lexical semantics. And yet, it still outperforms the basic model.

We were initially puzzled by the performance of the basic model after using this word information. (In Table 3.23, only a marginal 2.2% increase is observed to be contributed by this factor.) On the one hand, this seems counter intuitive. One would intuitively expect word meaning to play an important role (at least more important than 2.2%) in binding a pronoun to its antecedent. On the other hand, this statistic is collected from a much larger data set (1-million-word) than the coreference-marked training corpus on which the gender statistic $P(\rho \mid a, W_a)$ is collected, and yet contributes much less than the gender statistic which Table 3.23 shows to increase the accuracy by 12.5%. We hypothesized that maybe the data set is still not large enough since $P(w \mid h, t, l)$ is indeed more complicated than other statistics. We first need to find out if this hypothesis has any bases at all. To this end, we examine the relationship between the frequency of the head ***h*** and the accuracy with which the pronoun under it is resolved. We also examine the relationship between the frequency of the head noun of the antecedent and the accuracy. We plot these relations and the plots are shown in Figure 3.15 and Figure 3.16.

In Figure 3.15, the x-axis represents the number of times the head ***h*** is observed and they are bucketed. The y-axis is the probability reflecting the resolution accuracy for the pronoun under that head. The upper curve is for correctly resolved pronouns, i.e. $P(\rho \text{ correctly resolved} \mid count(head))$ and the lower curve are the cases where the resolution is wrong, i.e. $P(\rho \text{ incorrectly resolved} \mid count(head))$. Figure (3.16) does the same analysis using the frequencies of the head noun. Despite the fluctuations, a small trend can be observed. The

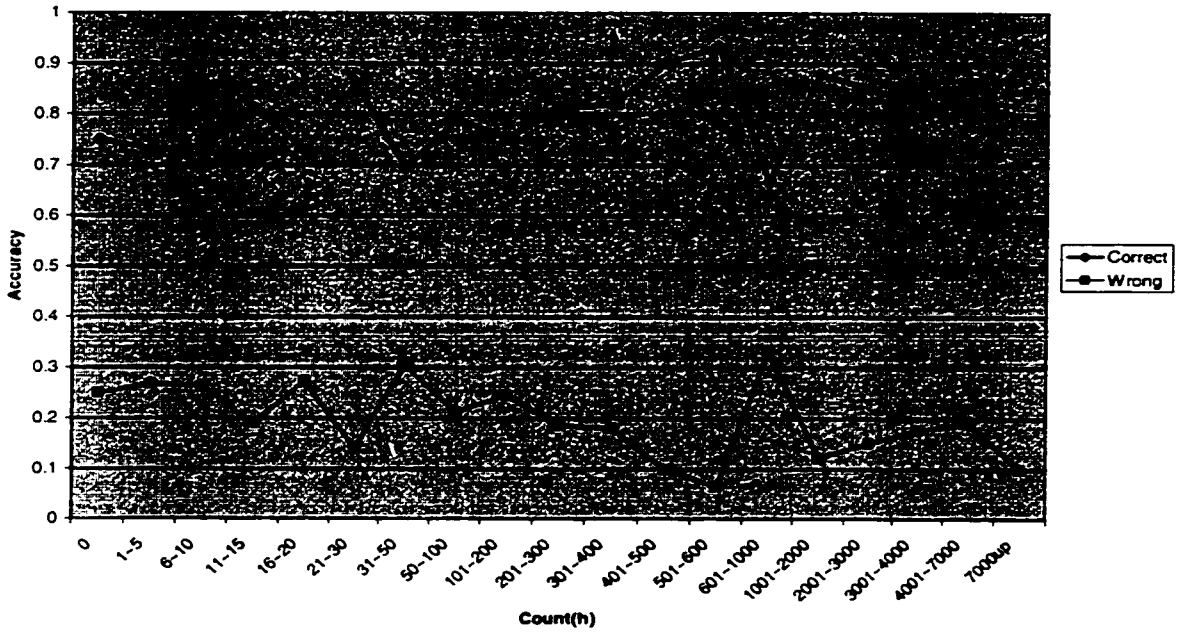


Figure 3.15 Accuracy vs. Head frequency

accuracy goes up ever so slightly with how well the head or the head noun is known.

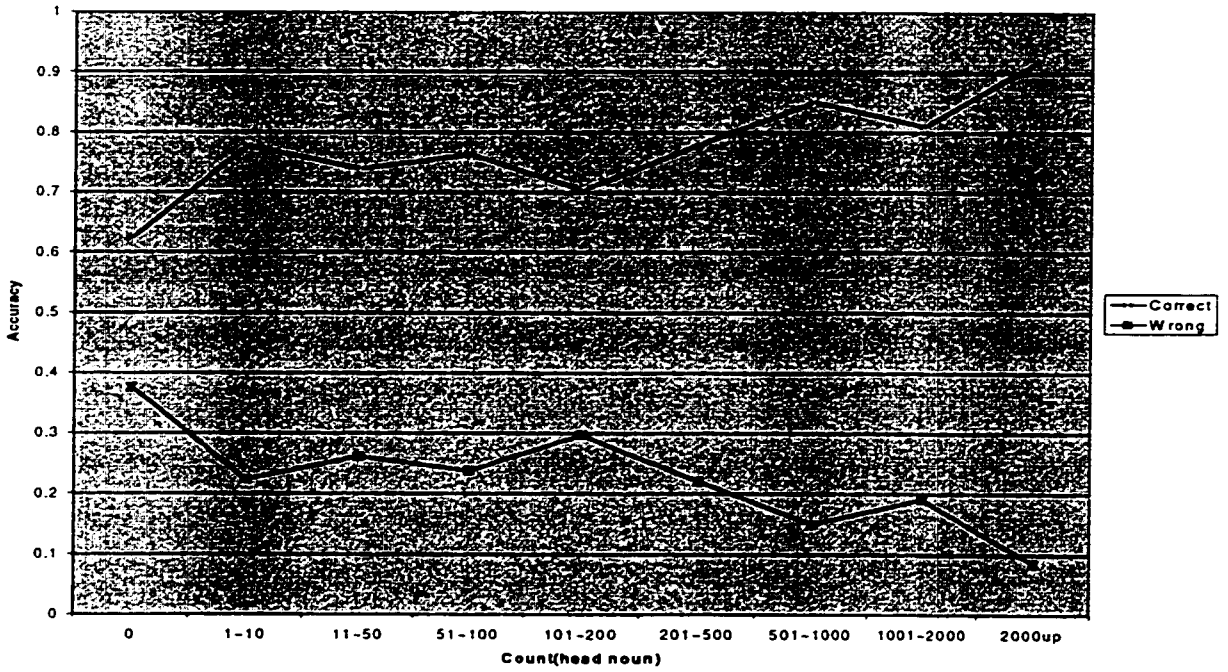


Figure 3.16 Accuracy vs. Head Noun frequency

Conversely, the inaccuracy curves go down with head (or head noun) counts increasing. These evidence led us to an experiment of collecting $P(w \mid h, t, l)$ on a much larger corpus (15 times bigger than the original one) in the hope of alleviate the sparse data problem. After gathering this statistic from a 15-million-word corpus and using it in the basic model, we see no improvement in the resolution accuracy. For this newly gathered statistic, a similar analysis of the relationship between frequencies of head/head noun and the resolution accuracy is carried out. Figure 3.17 shows the graph for the head noun analysis. The same trend is again present but does not seem to grow/drop faster. Figure 3.18 puts (3.16) and (3.17) together. The rate of growth (or decrease) does appear to be different but the change is just too tiny to have any statistically significant impact.

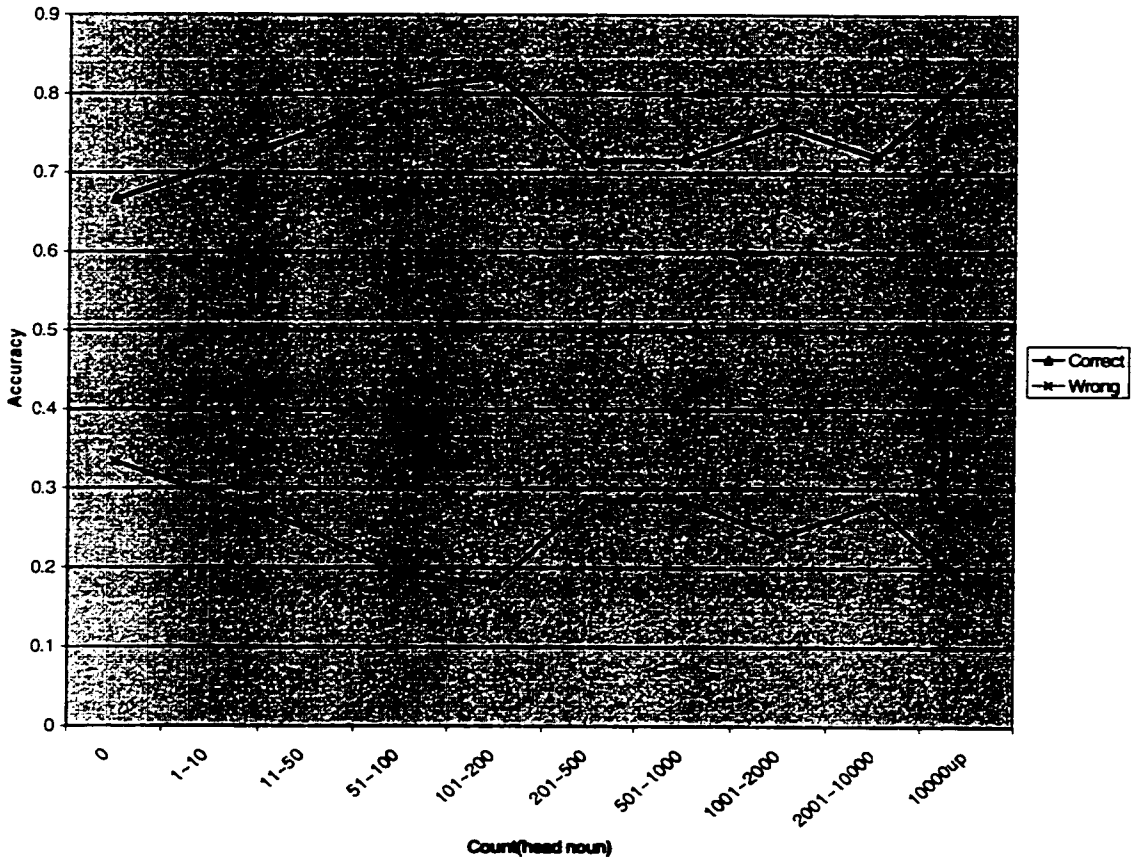


Figure 3.17 Accuracy vs. Head Noun frequency: large corpus

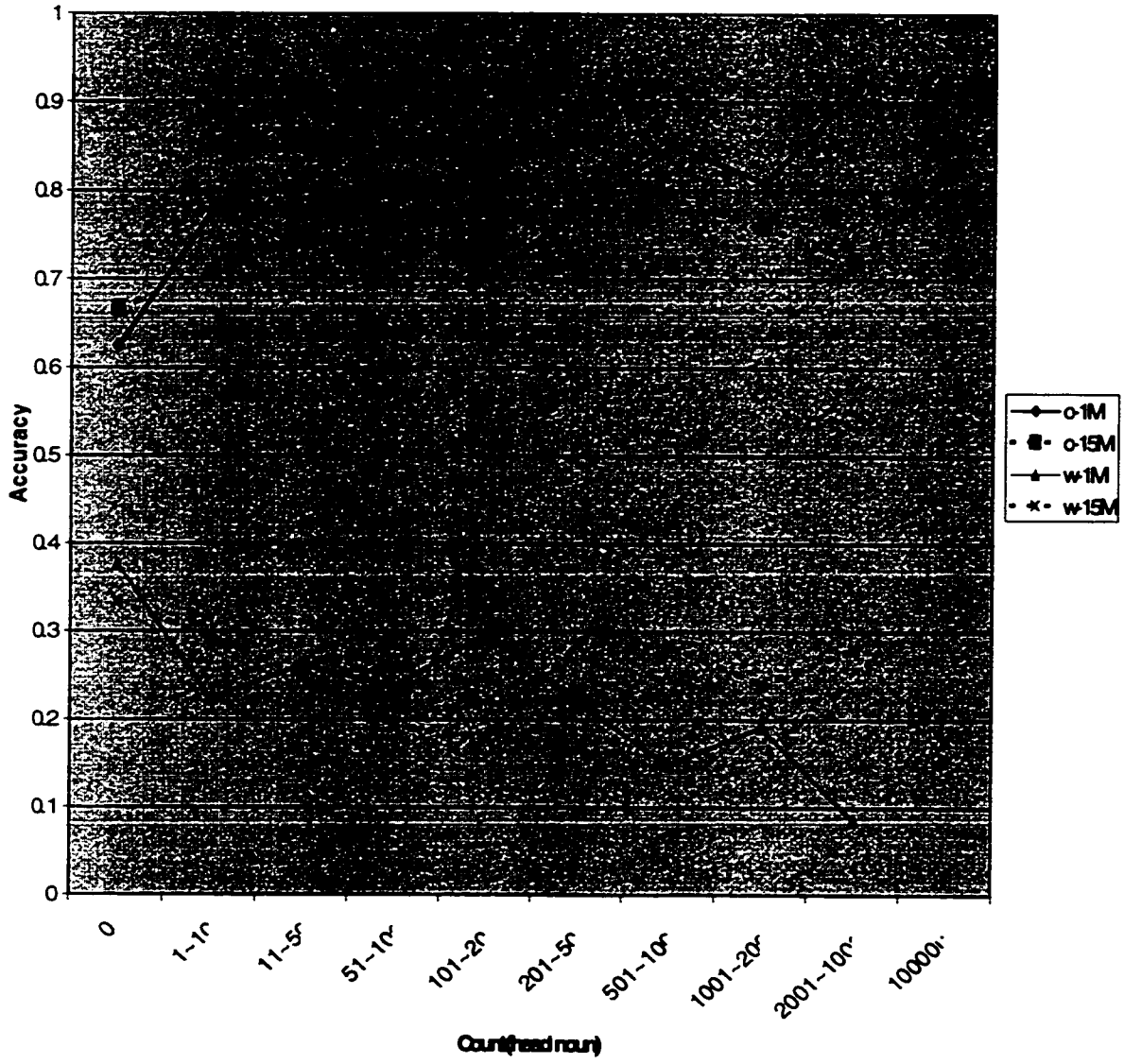


Figure 3.18 Accuracy vs. Head Noun frequency: small and big corpus

Since the head tuple *(h, t, l)* can be assumed to be independent of the other context parameters in the syntactic-prominence model, the $\frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)}$ term can be readily combined with equation (3.46). Experimental results show a 0.2% improvement by the model incorporating

the lexical semantics, hardly statistically significant. We speculate that the 2.2% increase in the basic model could very well be noise and the 0.2% in the syntactic-prominence model is simply a reflection of that noise.

We thus decided not to use the lexical semantic factor in the syntactic prominence model. What this model uses which is absent in the basic model is $P(G_{wa} \mid a, G_p)$, the grammatical role relationship between the pronoun and its antecedent. From Table 3.24, it is clear that this factor is what makes this model outperform the basic model. As discussed in section §2.5, the factor of syntactic prominence is used by all of the four algorithms and it is not surprising that putting it in our statistical framework improves the resolution accuracy. Table 3.4 shows that antecedents do tend to occupy certain syntactic roles with greater probability than other roles. Researchers have observed the special role of subject for a long time. Subject position noun phrases usually correspond to themes, topic, proto-Agent, sentence-aboutness, etc. It is regarded as the most salient position in a sentence. As for other positions, the ordering of prominence is less straightforward. For example, in centering theory, there is generally no agreed-upon ordering of the Cf list (Cote 1998, Strube & Hahn 1999) other than the subject. In almost all the orders proposed in the literature, the subject is always the highest ranked element. It is a common feeling that the outrank by subjects is the greatest while others often vary (such as object > adjuncts). The statistics we collected (Table 3.4) show one way of quantifying the relative salience of different grammatical roles.

3.4.4 Classifying errors

One way of comparing different factors in anaphora resolution is by running incremental experiments (see section §3.4.2). Another way is by classifying the errors. This will tell us something about each factor's potential as well as what is missing in the system. According to the factors we use in the system, we distinguish four types of errors and they are:

- ***Miss***: The correct antecedent is not even in the candidate list. This can happen to some cataphora cases and pronouns whose antecedents are at Hobbs distance greater than 25.
- ***Gender/Number/Animacy***: The proposed antecedent disagrees in gender/number/animacy feature(s) with the pronoun. In other words, if this information were perfect in the system, the correct antecedent would have been selected.
- ***Lexical semantics***: The proposed antecedent is very unlikely to occur under the head of the pronoun, i.e. the answer would have been correct if we had perfect lexical semantic information.
- ***World knowledge/Context-based inference***: The parameters in the system, no matter how perfect they are, are simply not enough to resolve those pronouns.

The analysis shown in this section is done on the output of the syntactic prominence model and is done for two classes of pronouns: the HE/SHE/IT class and the THEY class pronouns[1]. The percentages of each type of errors are shown in Table 3.25.

Error Type	HE/SHE/IT	THEY
Miss	7.7%	5.9%
Gender/Number/Animacy	21.2%	15.7%
Lexical semantics	19.2%	17.6%
World knowledge/Context-based inference	51.9%	60.8%

Table 3.25 Error classification

Since we use Hobbs' algorithm as a way to collect candidates, our system will fail to find the correct antecedent if Hobbs' tree walk misses it or if the antecedent occurs far back from the pronoun. An example of the former case is shown by sentence (3.14):

[1] The I/WE class pronouns are relatively scarce and this kind of analysis is not very informative.

3.14 In ***its*** latest compilation of performance statistics, ***Moodys' Investors Service*** found that investment-grade bonds posted a total return of 2.7% in October.

Cases of this nature (cataphora in general) may need special treatment. Cases where the antecedent is far away generally cannot be solved by simply increasing the candidate list size. Increasing the size will only include them in the list. The probability of them being the correct antecedent given their large distance is close to 0 (observe how fast Hobbs distance probabilities drop in Table (3.10)). Unless all the candidates preceding it are absolutely ruled out (which is highly unlikely), those cases appear "*hopeless*" to the system.

Because all information is expressed in a probabilistic way and no manual testing is involved, the gender/number/animacy statistic is far from being perfect, even with all the additional helpers. Errors of this nature are inevitable. An example is shown in sentence (3.15):

3.15 Outside, a young pressman filling a news box with an extra edition headlined "Herald Examiner Closes" refused to take ***a reader***'s quarter.
"Forget it," he said as he handed ***her*** a paper.

"*a news box*" occurs at Hobbs distance 4 which makes it 2.16 times as likely as "*a reader*" which occurs at Hobbs distance 6. The probability for "*reader*" being a SHE, i.e. *P(SHE | reader)* is greater than that for "*box*" *P(SHE | box)*, but only by 1.19. The combined probability turned out to be higher for "*a news box*" than for "*a reader*". Ideally, the probability for "*box*" should be close to 0[1]. In the HE/SHE/IT column of Table 3.25, 72.7% of the gender/number/animacy errors occur to HE/SHE and the remaining 27.3% are for IT. This means HE/SHE class pronouns would have the most gain from such information being perfect. This is not only because there are fewer HE/SHEs than ITs in the Wall Street Journal, but also because those pronouns have more restrictive requirement in the compatibility of their antecedents. The antecedents need to be animate as well as being masculine for HE and feminine for SHE whereas almost any inanimate object can be an IT.

---

[1] The word "*box*" does not occur in the WHICH class either. Otherwise it would have been ruled out.

In the THEY class, these errors are due to number disagreement. The naïve heuristic (section §3.3.4) for ruling out non-collective singular nouns apply only to singular non-proper name nouns. Companies are very likely to be referred to simply by their names and subsequently pronominalized by a THEY pronoun. For this reason we cannot rule out singular proper names. In the following sentence

3.16 *Charles Wohlstetter* has received countless letters from ***other CEOs*** offering support.
"***They*** said universally, without a single exception. Don't compromise."

"*Charles Wohlstetter*" is a proper name. Although it is singular it is not ruled out as being possible for the "*they*" in the second sentence of (3.16). The ideal statistic would tell that "*Charles Wohlstetter*" is probably a single person's name and its probability of being coreferred with a THEY-pronoun is very small. Presently, there are no techniques employed to distinguish different kinds of proper names. Table 3.25 also indicates that the ideal gender/number/animacy information would help HE/SHE/IT pronouns more than the THEY pronouns. This is because the THEY class pronouns are not only exempted from number test, they are also exempted from animacy test. Animate as well as inanimate objects can be THEY.

Table 3.25 shows that a good lexical semantics could reduce the error rate by a little less than 20%. Such information would help in sentence (3.17):

3.17 What saved ***many farmers*** from a bad year was the opportunity to reclaim large quantities of grain and other crops that ***they*** had mortgaged to the government under price support load programs.

The incorrect antecedent for "*they*" proposed by the system is "*grain and other crops*". Given that the head is "*mortgaged*", the possibility for "*grain and other crops*" should become very small. Ideally we would like to see P(farmers | mortgage, S, NP) >> P(crops | mortgage, S, NP).

Over half of the errors are due to the lack of world knowledge or context-based inference mechanism. If given the tuple ***(pronoun, its head, correct antecedent*** A_c, ***proposed antecedent*** A_p***)***, it is still not clear why A_c is the antecedent while A_p is not, then the error belongs to this category. Consider the tuple *(it, (demonstrates, S), the finding, the AT&T team)*. This tuple

depicts a situation where the pronoun "*it*" is the subject of "*demonstrates*", the correct antecedent is "*the finding*" and the antecedent proposed by the system is "*the AT&T team*". For the author at least, based solely on these information, both antecedents seem plausible. It is not clear why "*the finding*" is the intended antecedent. It all became rather obvious once the actual discourse were processed:

3.18 The AT&T team created the desired crystal changes by bombarding superconductor samples with neutrons.
Still, scientists breathed a collective sigh of relief about ***the finding*** because ***it*** demonstrates how to overcome the "flux pinning" problem.

Although "*the AT&T team*" occurs in one sentence before the pronoun, its being in unmarked subject position gives it much higher probability than the closer one "*the finding*".

A more striking example is the tuple *(his, (departure, NP), Mr. Reupke, Mr. Sheppard)*. Here the pronoun is a possessive "*his*" and it is followed by a noun "*departure*", "*Mr. Reupke*" is the correct antecedent and the program picks out "*Mr. Sheppard*". In terms of gender or lexical semantics, both candidates are equally good. We need the actual discourse context to disambiguate:

3.19 Reuters Holdings PLC said ***Mr. Reupke*** resigned as general manager to pursue unspecified interests.
Mr. Sheppard, an analyst at UBS Phillips & Drew in London, said, "I suspect the departure will be fairly irrelevant for the company."
Reuters said ***his*** departure reflects "no change in strategy or profits."

After learning from the context and acquiring the knowledge that it was *Mr. Reupke* who was leaving and *Mr. Sheppard* was just commenting, it became quite clear which one is the correct antecedent.

For cases like these merely improving the statistics is not going to help a lot. In fact, for cases like (3.19), it does not seem that the resolution factors currently employed by the system, no

matter how good their statistics get, will not be able to handle them correctly and for the correct reason[1].

We noted above that the THEY-class pronouns are more difficult partly because they do not encode clear gender/number/animacy information as the HE/SHE/IT pronouns do. Consequently one would expect to have more need of world knowledge and context modeling to correctly resolve them. This can be seen in the last row of Table (3.25). A larger percentage of the THEY pronouns (60.8%) are in this category than the HE/SHE/IT pronouns.

Before closing this section, there is an interesting finding from this error analysis that is worth mentioning. For the cases involving intersentential antecedents, we are interested in knowing if the centering algorithm would get them right. To meet the requirements of the centering algorithm, we exclude the following cases:

- the correct antecedent is intersentential but our system picks an intrasentential one. The centering algorithm applies to intersentential antecedents. Our evaluation shows that if the system were told to look for only intersentential antecedents for these case, it would have got the answers correct.
- the correct antecedent is at two sentences or further back from the sentence where the pronoun occurs. It is not clear how the centering algorithm will handle these cases and therefore they are excluded from this evaluation.
- our program picks a wrong intersentential antecedent (in the immediate previous sentence) due to wrong gender/number/animacy information. In other words, had we had perfect gender/number/animacy knowledge like the centering algorithm assumed, we would have got the correct answer.

The cases we apply this evaluation to are therefore those where the correct antecedent and the incorrectly proposed antecedent by the program both appear in the immediate preceding sentence

---

[1] It is possible that some of these pronoun can get resolved correctly by improved statistics but not for the

and they both agree in gender/number/animacy feature with the pronoun in question. The result is in Table 3.26.

Result	Count	Percentage
correct	1	16.7%
wrong	5	83.3%

Table 3.26 Evaluation of the centering algorithm on the incorrect output

The percentages may be a little overdramatic due to the fact that there are only 6 such cases under evaluation. The case of where the centering algorithm gets the correct answer is given in sentence (3.20). This is a discourse segment centered on the *insurance company*.

3.20a Wednesday's dominant issue was ***Yasuda Fire and Marine Insurance*** which continue to surge on rumors of speculative buying.
b ***It*** ended the day up 80 yen.

The two competing candidates are "*Wednesday's dominate issue*" (henceforth *issue*) and "*Yasuda Fire and Marine Insurance*" (henceforth *Yasuda*). They result in the following transitions:

Candidate	Cb(3.20b)	Cp(3.20b)	Transition
It = "issue"	issue	issue	SMOOTH-SHIFT
It ="Yasuda"	Yasuda	Yasuda	CONTINUE

Table 3.27 Centering algorithm applied to error output:correct

CONTINUE is preferred to SMOOTH-SHIFT and thus *Yasuda* is chosen as the antecedent. In our program, the antecedent *issue*, being in the subject position, was preferred.

An example where the centering algorithm offers no help is given in (3.21). This is a discourse centered on the city of Los Angeles.

3.21a ***The Los Angeles*** Times, with a circulation of more than 1.1 million dominates the region.
Cb = {the region: Los Angeles}
b But ***it*** faces stiff competition in Orange County.

---

right reason, i.e. not by context inferences, but most likely by chance.

The "*the region*" in (3.21a) refers to *Los Angeles*. Our system incorrectly picks out "*the region*" as the antecedent for the "*it*" in (3.21b)[1]. The centering approach won't get the correct answer either as can be seen from Table 3.28:

Candidate	Cb(3.21b)	Cp(3.21b)	Transition
it = "The Los Angeles Times"	the Los Angeles Times	The Los Angeles Times	SMOOTH-SHIFT
it = "the region"	the region	the region	CONTINUE

Table 3.28 Centering algorithm applied to error output: incorrect

"*the region*" results in a CONTINUE transition and is preferred to the SMOOTH-SHIFT by "*The Los Angeles Times*". The centering algorithm will still propose "*the region*" as its antecedent.

A sample of intersentential antecedents which are correctly resolved by our system is randomly drawn[1] and the centering algorithm is applied to them. The result is shown in Table 3.29.

Result	Percentage
correct	60%
wrong	40%

Table 3.29 Centering algorithm applied to the correct output

Although it is clear from Table (3.26) and (3.29) that applying the centering algorithm will result in a net loss in the resolution accuracy, this is not meant to discredit the centering theory. What this indicates is that our current system has incorporated the computational elements that the centering algorithm uses. It shows that we have combined the factors used in the centering algorithm in our statistical framework, most notably the grammatical salience and the discourse centers (approximated by the mention counts). The BFP centering algorithm, being a deterministic one, rules on absolute basis whereas our statistical approach can readily combine other resolution factors. The failure of the centering algorithm on the cases analyzed does not

---

[1] Although "*The Los Angeles times*" occupies a subject position (thus is generally preferred), it has lower mention counts (first time being mentioned) than "*the region*" (the 5th time). The combined probability for "*the region*" comes out higher.

mean failure of the centering approach. Rather, it indicates that there are more linguistic evidence and contextual parameters than what are present in our approach and in the centering approach.

Examples like (3.21) also show a general difficulty with the centering algorithm. Although it uses a grammatical role hierarchy to order the Cf list, it does not use this ordering to influence antecedent selection. According to Constraint 3, the centering algorithm does not require that the highest ranked element of $Cf(U_{i-1})$ (for example, "*The Los Angeles Times*" in 3.21a) actually be "*realized*" in U_i, only that $Cb(U_i)$ be the highest ranked element of $Cf(U_{i-1})$ which is in fact realized in U_i.

3.5 Comparison with previous approaches

In this section I compare our approach with the previous approaches described in Chapter 2. The comparison is not as straightforward as it may first seem. Different algorithms make different assumptions, and require different data input. The evaluation of some were manual and of others were automatic (i.e. computerized) The data domains differ, the sizes of the test data differ, the ranges of pronouns different algorithms target differ, and so on. Given this variety of differences, it is very hard, if not impossible, to come up with a single all-encompassing criterion according to which each algorithm can be rated. Instead, I draw up a table of the differences and compare performance in the common areas. The comparison is shown in Table 3.30.

It is pointless to compare an automatic program to a manual evaluation. Since the general idea of Natural Language Processing is to process languages on computer, the algorithms that can be implemented in a computer program seem to have the advantage in this regard.. Among the five algorithms, the RAP system and our system are computerized. The common features of the two approaches make some comparison possible. Our accuracy on anaphoric

---

[1] We need to make sure that there is more than one intersentential candidates that are compatible with the pronoun in gender/number/animacy feature. If there is only one, the centering algorithm will get it correct

HE/SHE/IT pronouns is certainly very high (91.1%). But the accuracy for pleonastic ITs is rather poor (for reasons, see section §3.4.2). We tried to devise more patterns. That got more pleonastic ITs correct but the precision started dropping, lowering the accuracy for anaphoric pronouns. It seems that patterns for pleonastic recognition involve both syntactic and lexical considerations. For example, the adjective pattern is a syntactic constraint but is also influenced by the actual adjective used in the sentence. That gives us the idea of collecting a mini adjective dictionary. In general, it is not clear how to devise a probabilistic pleonastic pattern matching method that combines the generality of syntax and the specificity of lexical choice while avoiding the sparse data problem. The approach in RAP is a deterministic rule-based method. Since they have got all the possible adjectives, passive verbs, and so on in the data, their recognition is perfect for pleonastic ITs in their test data. Note however, the set of rules identified in RAP will not cover those cases in §3.4.2 reproduced here in sentences (3.22) to (3.24)

3.22 ***It*** could take years for the new Polish government to fully use the aid effectively.
3.23 The House passed legislation designed to make ***it*** easier to block airline leveraged buy-outs.
3.24 Usually ***it*** is large investors initiating a buy or sell in Chicago.

There are no rules that account for these sentences in RAP (obviously, these cases did not occur in RAP's data). In addition to the deterministic rules, RAP also shifts some of the burden to its parser. The English Slot Grammar parser used by RAP does, in fact, recognize some pleonastic uses of "*it*", especially in constructions involving extraposed sentential subjects, as in:

3.25 ***It*** surprised me that he was there.

A special slot for the "*it*" is used. This is equivalent to the use of EXPLETIVE empty nodes in our parse trees. However we decided that it was the job of pronoun resolution to recognize them (they are after all, pronouns) and therefore we chose not to use the EXPLETIVE empty nodes.

---

because it has "perfect" gender knowledge.

Feature	Hobbs	BFP Centering	RAP	Mitkov	Ge&Charniak
Automatic	No	No	**Yes**	No	**Yes**
Coverage (type)	anaphoric	anaphoric	anaphoric + pleonastic	anaphoric	anaphoric + pleonastic
Coverage (pronoun)	HE/SHE/IT /THEY	—	HE/SHE/ IT /THEY	IT	All[1]
Test data size (number of pronouns)	100 (300)[2]	94 (281)	**360**	56	207 (**2069**)
Data restriction	No	intersentential	**Filtered**[3]	No	**No**
Data format	Parsed	Parsed	Parsed	POS tagged	Parsed
Perfect gender/number/animacy	Yes	Yes	**Yes**	Yes	**No**
Selectional restriction	Yes	No	No	No	No
Performance (HE/SHE/IT)	89.9%	—	—[4]	89.7%	**92.2%**
Performance (anaphoric)	91.7%	76.5%[1]	**85.3%**	89.7%	**88.5%**
Performance (pleonastic)	—	—	100%	—	54.9%
Performance (overall)	91.7%	76.5%	86%	89.7%	87.6%

Table 3.30 Comparison of the algorithms

[1] We resolve *All* anaphoric and pleonastic pronouns. The *YOU* pronouns are deictic and are not evaluated.

[2] The first number is the average number of pronouns and the number in parentheses is the total number.

[3] See section §2.3 for detail.

[4] There is no accuracy reported for HE/SHE/IT in the RAP paper (Lappin & Leass 1994).

Pleonastic ITs, important as they may be, are in general rare cases. There are roughly 6% of such cases in RAP's test data and in our test data only 2.6%. In a pronoun resolution system, they do not seen to be on the top of priorities. In the anaphoric domain, our system performs about 2% better than RAP. The RAP evaluation does not give accuracy for HE/SHE/IT and therefore cannot be compared with our 91.1%. The overall performances of the two systems are very competitive. This is not surprising since the factors used in both systems are very similar: gender/number/animacy, syntactic constraints, grammatical roles, frequency of mention, etc. RAP certainly got much better gender/animacy information than we did. If we had got the same good knowledge, we would have got the correct antecedents in the cases where the errors are due to the lack of such knowledge (21.2% among HE/SHE/IT and 15.7% among THEY. See section §3.4.4). Thus with perfect gender/animacy information, our performance for anaphoric pronouns would have been 89.3% and the overall accuracy would have been 88.1%. It can also be seen from Table 3.30 that our system has a wider coverage than RAP and imposes no restriction on the test data.

---

[1] This result is from Walker's evaluation (Walker 1989).

Chapter 4 Factors In Anaphora Resolution

Now that we have seen five approaches to the anaphora problem, in this chapter I would like to draw some conclusions regarding various factors in anaphora resolution. Researchers in formal linguistics and NLP alike have wondered "what do we really need in anaphora resolution and how much?" These are difficult and deep questions. It is not my intention to give decisive answers to these questions. Rather, this chapter is a discussion on such matters based on some empirical evidence presented in previous chapters.

4.1 Pronouns themselves

It is generally agreed upon that pronouns by themselves have very little syntactic or semantic content. Pronoun encodes features such as number, gender, animacy, reflexitivity but little, if any, other semantic content. That little content, however, turns out to be very important. In section §3.4.2., we saw that Hobbs' algorithm without this information performs 67.7% accurate and jumps to 80.2% with this information. In RAP a similar evaluation of Hobbs' algorithm is also given, assuming perfect knowledge and their experiment showed an accuracy of 82%. In centering, it has been suggested (Kehler 1993) that the extent to which SHIFTs cause additional processing load depends on the agreement features of pronouns and their antecedents. GJW (GJW 1995) notes that in example (4.1), the reference in sentence (4.1e) causes the reader to be misled:

4.1a ***Terry*** really goofs sometimes.
b Yesterday was a beautiful day and ***he*** was excited about trying out his new sailboat.
c ***He*** wanted Tony to join ***him*** on a sailing expedition.
d ***He*** called him at 6am.
e He was sick and furious at being woken up so early.

According to the BFP algorithm, in (4.1e), "*he*" coreferring with "*Tony*" constitutes a SMOOTH-SHIFT transition whereas coreferring ith "*Terry*" constitutes a CONTINUE relation. This correctly predicts the oddness of this passage. If this example is modified so as to make *Terry*

female, then there is little difficulty in interpretation despite the SMOOTH-SHIFT in the final sentence:

4.2a ***Terry*** really goofs sometimes.
b Yesterday was a beautiful day and ***she*** was excited about trying out ***her*** new sailboat.
c ***She*** wanted Tony to join ***her*** on a sailing expedition.
d ***She*** called him at 6am.
e He was sick and furious at being woken up so early.

Unlike the garden path effect exhibited by (4.1e), the identical (4.2e) is easily processable since the pronoun has only one possible referent. Kehler (Kehler 1993) goes on to propose that as long as a noun phrase is coreferential with the most highly-ranked member of the preceding Cf list for which no gender/number/animacy constraints apply, then that noun phrase can be pronominalized.

The effectiveness of this factor is also present in RAP. In 34% of the cases that the algorithm resolves correctly, the morphological filter reduces the set of possible antecedents to a single NP. In Mitkov's evaluation he reported 89.7% success rate. The success rate is, however, only 82% for those pronouns which after activating the gender/number/animacy filters, still have more than one candidate for antecedent (Mitkov calls it the *critical success rate*). The importance of this feature can also be seen from the distribution of our errors among different pronouns. Among the HE/SHE/IT pronouns which are incorrectly resolved, 73.2% are ITs. This in part is because (at least in WSJ) there are more inanimate objects than human entities. Unlike HE/SHEs which further require the separation of male entities from females, the ITs can be any singular inanimate objects. Table 3.21 shows that the accuracy for THEY is much lower than that for HE/SHE/IT. One of the reasons is that the THEYs encode even less information. The best we can say about the THEYs is that they prefer plural nouns. They can certainly be either animate or inanimate. We have seen that they can also refer to singular nouns. Consequently, this makes them more difficult to resolve than the singulars. In our system, taking out the additional helpers that provide extra gender information would hurt the performance by roughly 4%.

Although we know this feature is important, to the author's knowledge, no approach has been proposed to automatically acquire it. In all the four approaches we examined, this knowledge is either assumed (by human evaluation) or encoded in a lexicon or manually drawn up for some domain-specific data. Needless to say, such approaches are not feasible in processing large texts coming from diverse domains. In our approach we gather such information automatically from a training corpus with pronoun coreference marked. Our experiments show that even with a very small training set, the program still performs very well. For a broader coverage of words, we provide an algorithm for learning such information unsupervised by applying the pronoun resolution program to a large set of texts. Experimental results show that naïve as this approach may seem, the information learned is helpful when fed back into the anaphora program. (see section §3.3.3.1). Although currently the program is applied to the Wall Street Journal corpus, it is not domain specific and can be readily applied to other parsed texts.

4.2 Grammatical salience

Human languages, unlike programming languages that are designed for computers, exhibit a significant amount of ambiguity. Yet we humans do not have much problem processing them. It is this interest in ambiguity resolution that has led computational linguists to consider reference and coreference as crucial to natural language processing. One principal focus in pronoun's reference has been how the context of a sentence influences the choice of an antecedent. One of the central questions concerns the relation between inference in coreference and inference in coherence in general. Some researchers (e.g. Hobbs 1979) have offered important attempts to pronoun resolution by semantic inferences. Hobbs (Hobbs 1976) tries to show how pronoun resolution "*happens*" in a total system for semantic analysis. The word "*happens*" is, he argues, appropriate because once everything else is done, pronoun resolution "comes free — it happens automatically". Thus according to Hobbs, coreference is in effect a byproduct of general inference mechanisms that are used to make text coherent. He then

proposed a system that accomplishes pronoun resolution by unstructured semantic inferences "from a database of world knowledge."

However, there are two facts that are not explained by purely content-based models of reference and coherence. It is observed by linguistic and cognitive experiments (Gordon & Hendrick 1997, GJW 1995) that the coherence of a discourse depends not only on semantic content but also on the forms of referring expressions. In other words, proper names, definite noun phrases, and pronouns are not equivalent in terms of their effect on coherence. An example of such difference is provided by (1.1) where the use of a proper name in (1.1c') make the passage quite odd and the substitution by a pronoun in (1.1c) makes the sentence much more acceptable. Secondly, hearers have tendencies to assign referents to pronouns before the rest of the sentence is processed[1]. In sentence (4.1e), hearers/readers tend to assign "*Terry*" to the subject "*he*" initially and the semantic content of the rest of the sentence forces them to change the interpretation to "*Tony*".

It is these kinds of facts that has led to the contrasting view that language processing must take advantage of the contextual structure of language, particular with regard to reference, in order to constrain processes of inference and make them computationally tractable (Grosz 1977). This is the major motivation for the development of centering theory (see introduction in section §1.1). One of the general focuses in the centering approach is the role of syntactic prominence in coreference. Recall that ranking in the Cf list has two important consequences. It affects the likelihood that an entity will be the backward center of the subsequent sentence which in turn constrains the interpretation of pronouns. Whether or not grammatical role salience has any cognitive effect on anaphora resolution is not what I intend to answer. Rather, I want to address the effect of this factor on coreference. Empirical evidence that prominence of this sort influences coreference comes from studies of judgements of the acceptability of coreference and

reading time studies (Gorden & Hendrick 1997). It has been generally observed that a syntactic prominent antecedent facilitates coreference. The statistical results in this study corroborate these facts. We see that antecedents tend to occupy certain syntactic roles (e.g. subject) with greater probability than other roles (see Tables 3.4, 3.5 and Appendix B for details). These statistics from the *Collector* (Figure 3.14) then tells the *Resolver* to favor reference to entities with syntactically prominent antecedents over reference to entities with nonprominent antecedents. Using this factor enables the model to perform 3% better than the basic model. The use of this factor is present to varying degrees in all of the four approaches described in Chapter 2.

4.3 Discourse salience

In this factor I am mostly concerned with topics in a discourse segment. It is obvious that topics are more salient than other entities and hence are good candidates for pronominalization. In our approach, we use mention counts to approximate the identification of topics. Tables (3.6) and (3.7) show that as a story develops (i.e. as S_ρ increases), frequently mentioned entities are more and more likely to be realized as pronouns. This is not surprising since the primary function of pronouns (for this matter, any reduced expressions) is to refer to things that have already been mentioned in a discourse. The more times a thing is mentioned, the more central or salient it becomes in a discourse model, and the less necessary it is to refer to it by a full noun phrase[2]. We capture this phenomenon by using $P(a \mid M_a, S_\rho)$ in the basic model and by $\frac{P(M_a \mid a, S_\rho, G_\rho)}{P(M_a)}$ in the syntactic prominence model. The RAP system also uses frequency of mention, although it is not clear how this factor interacts with other components in RAP[3]. The *Lexical reiteration* indicator

---

[1] The garden path effects happen when the semantic information in the rest of the sentence/discourse contradicts the initial assignment and causes the hearer/reader to backtrack.

[2] The primary function of proper names and other full expressions is considered to introduce new entities into a discourse model.(Kamp 1993)

[3] There is no explanation in RAP (Lappin & Leass 1994) on where the counts come from or how they are used. One guess by the author is that it may have something to do with the *equivalence classes*. Members

is Mitkov's approximation to identifying discourse topics. The idea of utilizing discourse topic is implicit in the centering approach. Although there is no explicit use of frequency of mention, the more frequently something is mentioned (i.e. realized), the more likely it is to be the backward-looking center and hence the more likely to be realized as pronouns.

Discourse salience is also related to *recency*, i.e. the "closeness" of the antecedent to the pronoun. The farther away a noun phrase is from a pronoun, the less salient it seems to that pronoun. In Tables (3.9) and (3.10), we see the probability of a noun phrase for being an antecedent drops rapidly with distance. In both RAP and Mitkov, sentence recency (or *referential distance* as Mitkov calls it) is an important salient factor and proximity is the criterion for breaking ties.

All the above factors have been shown to be particular important to the anaphora problem. The gender factor acts quite independently of the other two factors. The grammatical role and discourse salience are closely related to one another. The grammatically salient role "*subject*" is often identified with the discourse salient role of topic or theme. In the RAP system, when various elements of the salience weighting mechanism were deactivated individually, the deterioration of the overall success rate is relatively small. When all structural salience weighting is switched off, the effect is a significant 27% drop. This suggests that the salience factors operate in a complex and interdependent manner for anaphora resolution. In our system, the interaction between grammatical and discourse salience is reflected by letting the distance and mention counts depend on the grammatical role of the pronoun and the form of the pronoun.

4.4 c-command

As we have seen in section §1.1, *c-command* was proposed as a mechanism for explaining some complementary distributions for reflexive and non-reflexive pronouns. As a

---

of an equivalence class refer to the same entity and therefore the size of an equivalence class tells how many times that entity has occurred in the discourse.

how much the constraints proposed by c-command contribute to anaphora resolution. To see this, we ran an experiment with the c-command constraints turned off. It turns out that the performance drops by only 1.2%. *c-command* works beautifully for contrastive sentences like (4.3) and (4.4)

4.3 Chomsky adores ***himself***.
4.4 Chomsky adores ***him***.

These sentences are relatively simple and many difficulties have been shown to arise when applying c-command to more complex sentences. (Pollard & Sag 1992). For example, some anaphors seem to be exempt from the binding principle for reflexives where there is no c-commanding antecedent at all as illustrated by (4.5):

4.5 The fact that there is a picture of ***himself*** hanging in the post office is believed to be disturbing to ***Tom***.

"*Tom*" is the antecedent for "*himself*" but "*Tom*" does not c-command the reflexive "*himself*" as demanded by principle A. We speculate that in real texts like WSJ, simple clear-cut cases like (4.3) and (4.4) are rare and sentences are, in most cases, longer and more complex. Because of the many problems with c-command on complex sentences, it does not always work very well. This is not to say that anaphora resolution does not need syntactic binding constraints. Rather, the experimental result showed that its relevance is not as strong as we initially thought. Therefore we observe only a small decrease in accuracy without it. One can certainly imagine texts in some domain (for example, children's story) consist mostly of simple sentences and c-command constraints on those data may be more effective and more influential.

4.5 Lexical semantics

In section §3.4.3, I indicated that the statistical encoding of selectional restriction is absent from the syntactic-prominence model and that when used in the basic model improves performance by only 2.2%. This fact is also observed by RAP. Experiments were conducted with the addition of a component that contributes statistically modeled information concerning

semantic relations. This enhancement, called RAPSTAT, only marginally improves RAP's performance by 2%[1]. The statistically measured lexical preference is essentially selectional restriction. The small effect of this factor can also be found in the manual evaluation of Hobbs' algorithm (section §2.1), where the perfect selectional restriction contributes 3%.

I suggested one possible reason (section §3.4.3) why the statistically collected semantic restriction offers little help. Although the graphs of Figure 3.15 — 3.18 show that the accuracy of this statistic change very slowly with increasing amount of data, the current evidence cannot rule out the possibility of the sparse data problem. The plots (and the evidence shown by RAPSTAT) do indicate that lexical semantic statistics, given their current formulation, are unlikely to be successfully collected on a large amount of data that is practically obtainable. The evidence exhibited by Hobbs hand simulation is more puzzling. Even the "perfect" selectional restriction improves accuracy by only 3%. This of course may be domain dependent. In Hobbs evaluation, this information helps most in the text from a history book (7% improvement) and least on the text from Newsweek (1%). Our WSJ corpus is certainly more like Newsweek in style than a history book. This seems to suggest that lexical semantics may be heavily domain specific.

Another difficulty with this factor is that it is more time dependent than other factors. Word meanings change over time and in fact, they change faster than we would expect. In RAP, an example is given arguing in favor of using RAPSTAT. The example is given in (4.6).

4.6a The users you enroll may not necessarily be new to the system and may already have a user profile and a system distribution directory entry.
b. &ofc. checks for the existence of ***these objects*** and only creates ***them*** as necessary.

RAP selects "*users*" in (4.6a) as the antecedent for "*them*" in (4.6b) and the correct answer is "*these objects*". RAP suggests that selectional restrictions can help in this case since *(create*

---

[1] The RAPSTAT is actually a mixed model in that the statistics are applied selectively. RAPSTAT is used when

i. the difference in salience scores between two candidates C_1 ad C_2 does not exceed a parametrically specified threshold, and
ii. the statistical score of C_2 is significantly greater than that of C_1.

objects) is more plausible than *(create users)*. Twenty years ago, *(create users)* would indeed by very awkward. Today, at least to the author, the pair *(create users)* is much more acceptable presumably because networks are so common nowadays and "*users*" do get "*created*" rather often.

4.6 World knowledge/Context-based inference

This factor is so broad that it is difficult, if not impossible, to characterize what is meant by the phrase "world knowledge". In a trivial sense, everything is world knowledge. As Hobbs (1976) said "(Charniak 1972) demonstrated that in order to do pronoun resolution one had to be able to do everything else." We certainly think that world knowledge and content-based inference are very important. In the example given by (3.19), there does not seem to be a way to tell why "*Mr. Reupke*" is correct and "*Mr. Sheppard*" is wrong based on the information we currently utilize. One needs to read the entire segment to figure out the referent. What enables us as humans to correctly resolve the pronoun "*his*" in (3.19c) is the inference we drew from the context and the knowledge we acquired from the segment. At the current stage it is not clear how this can be adequately modeled, either symbolically or statistically. World knowledge/context-based inference is by all means critical to anaphora resolution (or to any serious natural language understanding system), but the fact that we do not know how to "compute" them makes their importance irrelevant. There are things we know how to characterize and more importantly, how to compute such as the gender factor, the grammatical salience factor, and so on. These factors may not be as deep as world knowledge but they are important not only because they play a role in anaphora resolution but also because we can compute them and write computer programs for them.

4.7 Summary

In this chapter I have attempted to list the factors in anaphora resolution. Evidence form our own experiments and those from the Hobbs', BFP, RAP, and Mitkov's approaches suggest that gender/number/animacy feature, salience of grammatical roles, and discourse salience are major players in this problem, whereas constraints like those proposed by c-command and constraints from lexical semantics do not seem to be as generally important. This in part can be attributed to the relative "stableness" of each factor. In section §4.4 and §4.5, I have argued that c-command and lexical semantics could very well be domain specific. On some domain data, they could be very influential. In the data we deal with (the Wall Street Journal) they turn out to be less crucial. I have also noted that lexical semantics is time dependent. Statistically collected semantic relations from one period may not reflect the semantic relations in another period. In contrast, the other three factors are rather "stable" in the sense they do not change as rapidly and they can be applied universally i.e. not language dependent, and they can be easily learned when they do change.

The information encoded in pronouns in different languages differ but they all contain gender/number/animacy content to varying degrees and these features do not tend to change over time. Grammatical roles are usually determined by the word order of a language. Given the linearity of English, the first unit of a sentence is usually the subject and most salient. Other languages have different word orders and their own ways of identifying subjects. Word order is also a rather stable aspect of a language. Discourse salience factor, or discourse topichood is about the "*centralness*" of an entity. One way an entity becomes central is by being repeatedly mentioned. This fact is language independent, making this factor capable of being applied cross-linguistically. In contrast, binding constraints seem to be language specific. Kuno in as early as 1972 (Kuno 1972) has suggested that for some language like Japanese, pronouns are better analyzed by a discourse model than by syntactic binding principles. Evidence seems to show that

the usefulness of lexical semantics in anaphora resolution is domain specific. Also, corpus-based analyses like ours show that accuracy in statistically measured semantic relations is difficult to obtain. In addition, semantic relations, compared to other factors, have a higher tendency to change over a period of time. World knowledge and/or context-base inference is certainly very important. But everyone knows that they are very hard. Over the years many attempts have been made to address the issue. Shank (Shank 1973) and others tried to model the inferencing mechanism in special domains of knowledge. Hobbs (Hobbs 1976) devised a system consisting of certain semantic operations to show how they would assist in making appropriate inferences. Despite all these attempts we are still not ready to program "world knowledge". A considerable amount of progress has been made in the last few decades but we still have a long journey to travel.

Chapter 5

Further Applications in the Statistical Framework

In this chapter I will discuss how to apply the statistical framework presented in Chapter 3 to two related anaphora problems. Although the results are currently not used in our anaphora resolution system, the experiment and formulation point to the possibility of some future research.

5.1 Pronouns in text generation

This issue concerns with how to make pronoun/full noun phrase choices in text or dialogue generation. The problem is of some interest because pronouns and full expressions have different information values and hence have different effect on coherence (see section §4.2). Gordon & Hendrick (Gordon & Hendrick 1998) find that coreference is highly acceptable in sentences where a name precedes a pronoun in a [Name-Pronoun] sequence such as (5.1).

5.1 ***Lisa*** visited ***her*** brother at college.

Coreference is considerably less acceptable in sentences containing repeated names in [Name-Name] sequence such as (5.2).

5.2 ***Lisa*** visited ***Lisa***'s brother at college.

In an intelligent text/dialogue generation system, one would want to avoid the repeated-name penalty resulted from sentences like (5.2). One way to make a pronoun/full noun choice is by studying the distribution of coreferential pronouns and full nouns based on context. In our statistical anaphora resolution system we have identified a set of discourse context some of which are useful in this task. They are the grammatical role, the mention counts, and a derivable parameter "*competitor*". When facing a choice between using a pronoun or a full noun to refer to an entity *e*, the *competitors* are those entities in the discourse that are in the same gender class as

e. To facilitate coherence, one must avoid confusion and ambiguity. Entities having the same gender/number/animacy feature require the use of the same pronoun. Intuitively, the greater the number of competitors, the greater the possibility that a choice on pronoun can cause ambiguity. The way I propose to formulate the task is the following:

Let

- ***c***: choice of full noun or pronoun to refer to an entity ***e***
- M_e: the mention counts of ***e*** at the point of generation
- ***comp***: number of competitors at the time of generation
- G_e: the grammatical role that ***e*** occupies

Then ***c*** can be treated s a random variable whose value is either *FullNP* or *Pronoun* and the conditional probability sought is equation (5.1)

$$P(c = \{\text{FullNP, Pronoun}\} \mid M_e, G_e, \text{comp}) \qquad (5.1)$$

The three conditioning events M_e, G_e, and *comp* can be taken to be independent of each other. Applying Bayes' formula and the independence relations:

$$P(c \mid M_e, G_e, \text{comp}) \qquad (5.2)$$

$$= \frac{P(C, M_e, G_e, \text{comp})}{P(M_e, G_e, \text{comp})} \qquad (5.3)$$

$$\alpha\, P(c, M_e, G_e, \text{comp}) \qquad (5.4)$$

$$= P(G_e \mid c, M_e, \text{comp})\, P(M_e \mid c, \text{comp})\, P(\text{comp} \mid c)\, P(c) \qquad (5.5)$$

$$= P(G_e \mid c)\, P(M_e \mid c)\, P(\text{comp} \mid c)\, P(c) \qquad (5.6)$$

A program is then written to compute equation (5.6) and is run on the training data to collect the four statistics in that equation. The program only considers those entities with mention counts 2 or larger for the simple reason that when an entity is first introduced (M_e=1), in all likelihood, it will be introduced by a full noun phrase. Results show that the general tendency is to use pronouns to refer to existing entities (this supports the DRT's theory of primary function of pronouns). This is shown by the prior of ***c***:

[P(c = Pronoun) = 0.58] > [P(c = FullNP) = 0.42]

The results also indicate that embedded subject is the favorite position for pronouns. This can be seen in Table 5.1 where if the choice is *Pronoun* (i.e. row 3) the probability of occupying the ESBJ position (column 3) is the greatest (= 0.30). Also, when a possessive relation is needed, possessive pronouns (whose grammatical role is OTHERS in the system) are much preferred than full noun phrases by ('s) construction. This can be seen from the last column of Table 5.1 where the probability for *Pronoun* (0.17) is much larger than that for *FullNP* (0.005). This supports the findings mentioned in the beginning of this section that example (5.2) is much less acceptable than example (5.1). The statistic $P(G_e \mid c)$ is given in Table 5.1.

c(hoice)	UMSBJ	ESBJ	NPSBJ	OBJ	PP	PPS	OTHER
FullNP	0.37	0.21	0.08	0.12	0.21	0.007	0.005
Pronoun	0.27	0.30	0.009	0.06	0.18	0.01	0.17

Table 5.1 $P(G_e \mid c)$

The relationships captured by the other two statistics $P(M_e \mid c)$ and $P(comp \mid c)$ can be more clearly seen from another angle, namely $P(c \mid M_e)$ and $P(c \mid comp)$. The first one $P(c \mid M_e)$ is given in Table 5.2 and is plotted in Figure 5.1 where mention counts M are bucketed as before.

c(hoice)	M = 2	M = 3	M = 4	M = 5	M = 6	M = 7	M = 8	M = 9
FullNP	0.335	0.442	0.449	0.486	0.510	0.482	0.439	0.307
Pronoun	0.665	0.558	0.551	0.514	0.490	0.518	0.561	0.693

Table 5.2 $P(c \mid M_e)$

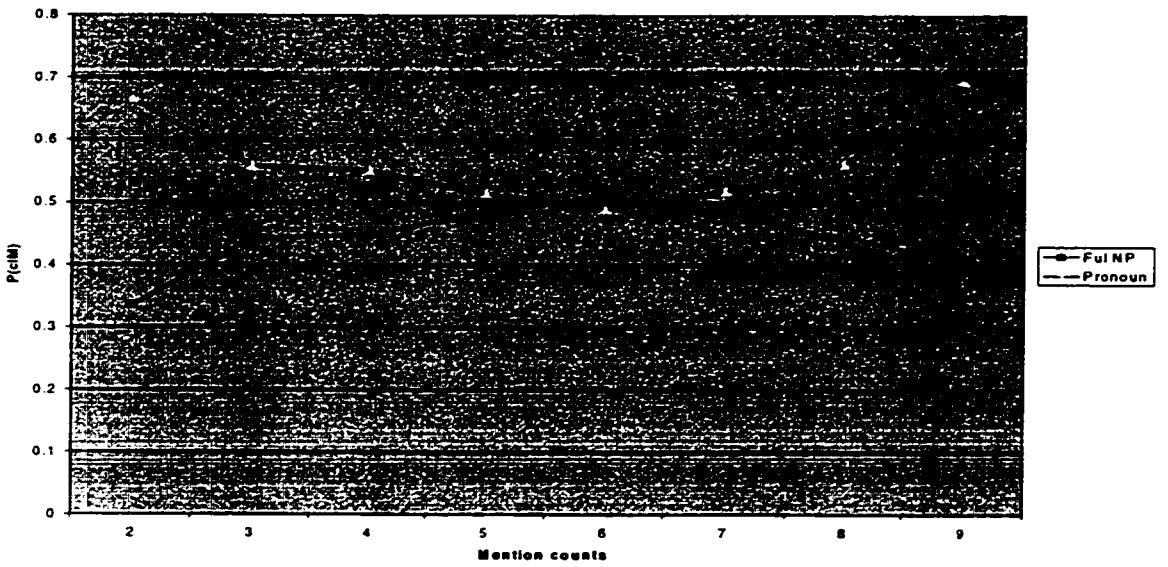


Figure 5.1 P(c | M_e)

These results show that when an entity is o be mentioned for the second time it is very likely to be realized as a pronoun. Such probability decreases for subsequence mentions (up to M=6) presumably because as the story develops more and more entities are introduced and that the range (M=3) to (M=6) is not quite sufficient enough for an entity to be identified as the topic, i.e. entities with mention counts [3,6] is not globally salient enough to license the use of a pronoun. As the story further develops and if an entity continues to be repeatedly mentioned, it becomes more and more globally salient, i.e. it gains more and more topichood. It is then salient enough to be realize as a pronoun.

The relationship between the number of competitors and the choice is shown in Table 5.3 and plotted in Figure 5.2. The number of competitors is not bucketed.

c(hoice)	comp = 0	comp = 1	comp = 2	comp = 3
Full NP	0.38	0.45	0.53	0.56
Pronoun	0.62	0.55	0.47	0.44

Table 5.3 P(c | comp)

The trends in Figure 5.2 are very clear. As the number of competitors increases, it becomes more and more necessary to use a full expression to avoid ambiguity. Conversely if the current entity is the only object (comp = 0) in its gender class then it is "*safe*" to pronominalize it.

In a text/dialogue generation program, one may use equation (5.6) to guide the choice between a pronoun and a full noun phrase. Intelligent uses of pronouns not only make a discourse more coherent and ease the processing of the discourse but also make the computer-generated language more "human-like".

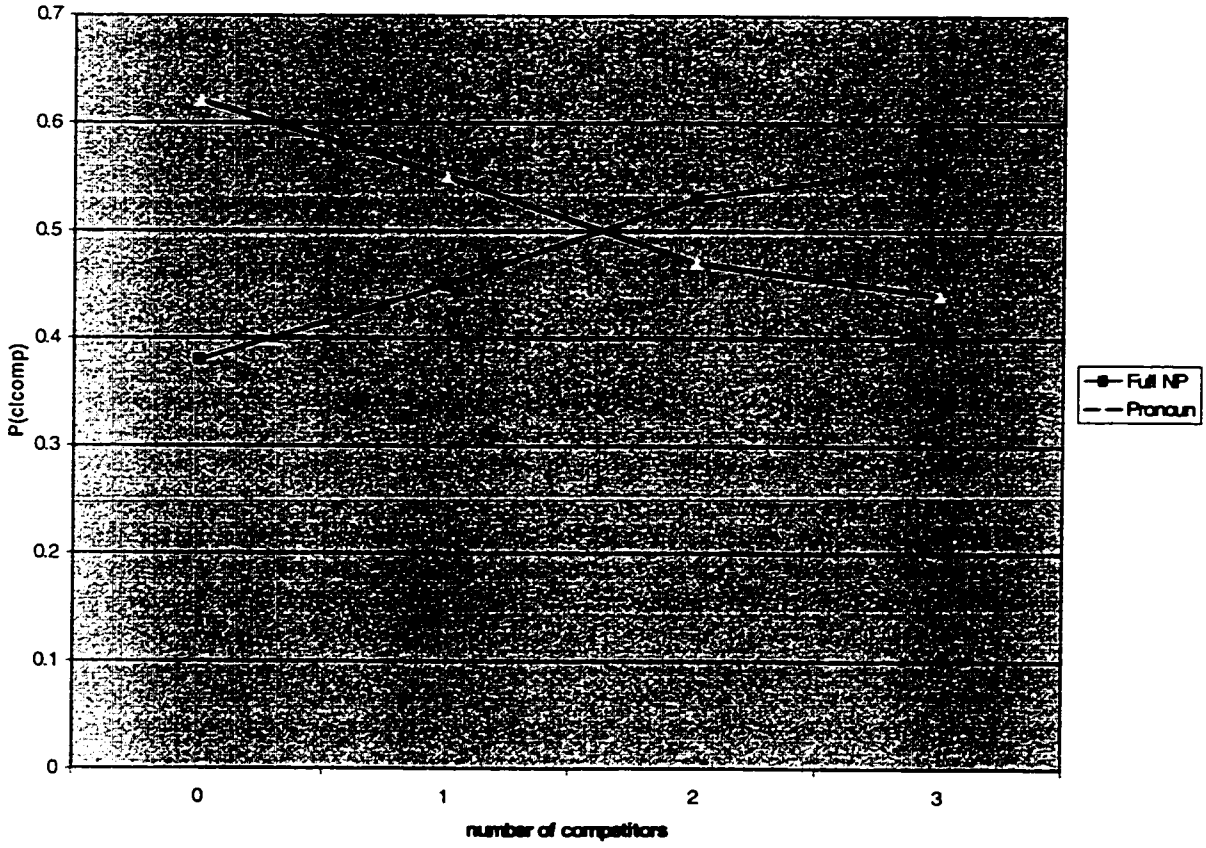


Figure 5.2 P(c | comp)

5.2 Centering revisited: a statistical attempt

In this section I propose a very naïve attempt to put centering in a statistical framework. Recall that the three data structures in centering are a list forward looking centers Cf, a backward-looking center Cb, and a preferred center Cp with the Cf list ordered according to the members' syntactic salience (see section §1.2). Coherence is then characterized in terms of transition from the $(i-1)^{st}$ utterance U_{i-1} to the i^{th} utterance U_i. The CONTINUE transition is considered the most coherent and hence preferred. In the statistical framework we then seek to maximize the probability of the CONTINUE transition. The naïve approach is as follows.

For a given pair of utterances U_{i-1} and U_i where U_i contains some pronoun ρ, we draw up the Cf list of U_{i-1}, denoted by $Cf_{i-1} = \{e_1, e_2, \ldots, e_k\}$ where the "e_i"s are the entities in U_{i-1}. Define the following events:

- Event A: $Cf_{i-1}[l] = A(\rho)$ for some $l = 1, \ldots, k$ where l is the index into the Cf_{i-1} list and $A(\rho)$ means *Antecedent* of ρ. This event says that one of the Cf_{i-1} members is the antecedent of ρ in U_i, henceforth written as $Cf_{i-1}[l] = \rho$.
- Event B: $Cf_{i-1}[l] = Cb_i$ for some $l = 1, \ldots, k$. Intuitively this is the event of one of the Cf_{i-1}'s members being the backward looking center of U_i.
- Event C: $Cb_{i-1} = Cb_i$, i.e. the backward looking center does not change from U_{i-1} to U_i.
- Event D: $Cb_i = Cp_i$, i.e. the backward looking center is the same as the preferred center in U_i.

We can assume that events C and D are independent of each other, but they do depend on events A and B. We then first calculate the probability of events A and B. Let

$$\alpha_l = P(AB) = P(A)\, P(B \mid A) \qquad (5.7)$$

In other words, for a particular index l, we calculate the prior probability that the l^{th} entity of Cf_{i-1} is pronominalized by ρ in U_i. If this is the case (i.e. conditioned on event A), then we would like

to compute the probability that it is the backward looking center of U_i (i.e. event B is conditioned on event A). Next let,

$$\beta_l = P(C \mid AB) \qquad (5.8)$$

After having assumed that $Cf_{i-1}[l]$ is the antecedent for ρ and is the backward looking center of U_i, we want to ask for the probability that it is the same as the backward looking center of U_{i-1}. Lastly, let

$$\gamma_l = P(D \mid AB) \qquad (5.9)$$

which calculates the probability that the backward looking center of U_i coincides with its preferred center. The CONTINUE transition occurs when both events C and D occur. For an assignment of antecedent of ρ, we would like to know the probability that this assignment results in a CONTINUE transition, i.e.

$$P(\text{CONTINUE}_l)$$

$$= P(A\ B\ C\ D) \qquad (5.10)$$

$$= P(A\ B)\ P(C\ D \mid A\ B) \qquad (5.11)$$

$$= P(A\ B)\ P(C \mid A\ B)\ P(D \mid A\ B) \qquad (5.12)$$

$$= \alpha_l\ \beta_l\ \gamma_l \qquad (5.13)$$

then the antecedent for ρ is the one that maximize the probability of resulting in a CONTINUE transition:

$$A(\rho) = \arg\max_l P(CONTINUE_l) \qquad (5.14)$$

$$= \arg\max_l \alpha_l \beta_l \gamma_l \qquad (5.15)$$

One difficulty that prevents equation (5.15) from being implemented on a computer is that in order to test some of the equivalencies (i.e. Events B and C), one would need to know if two full noun phrases corefer. Presently there has not been a high-accuracy automatic full noun phrase coreference resolution program. Also, given the relatively low accuracy of the centering

algorithm (76.5% see Table 3.30) on anaphora resolution, there are some doubts about how well centering theory can be applied to actual anaphora resolution program. The purpose of this attempt is not to actually build a statistical centering program. Rather, this is meant to suggest a general approach in this statistical framework.

Chapter 6 Conclusion

This concluding chapter has two sections. The first section gives a summary of the models. The second section suggests some topics for further research.

6.1 Summary of the models

The present approach to the anaphora problem took as its point of departure the restrictive and/or manual treatment of this phenomenon in previous works. Although RAP is a computer program, it operates on a manually filtered data set and perfect gender/number/animacy knowledge is at its disposal.

One of the goals of this research is to have a completely automatic anaphora resolution system that covers a wide range of pronouns. In fact, our system deals with all anaphoric pronouns. We choose to approach the problem by statistical means. The antecedent of a pronoun is treated as a random variable. The value of this random variable is the one that maximizes its probability in a given context. We devise two models corresponding to two sets of contexts. The **basic model** takes as its context the pronoun ρ in question, the candidate antecedents $\overline{W}$, the Hobbs' distances $\overline{d_H}$, the mention counts $\overline{M}$, the form of the pronoun f_ρ, the sentence in which ρ occurs S_ρ, the head environment surrounding ρ (h, t, l), and the *pattern* of the current sentence for recognition of pleonastic ITs. This basic model is formally written as

$$F(\rho) = \arg\max_a P(A(\rho) = a \mid \rho, \vec{d}_H, \vec{W}, h, t, l, \vec{M}, S_\rho, f_\rho, pattern) \quad (6.1)$$

$$= \arg\max_a P(d_a \mid a, f_\rho) P(\rho \mid a, W_a) \frac{P(W_a \mid a, h, t, l)}{P(W_a \mid t)} P(a \mid M_a, S_\rho) P(pattern \mid a) \quad (6.2)$$

This model incorporates the distance factor, the gender/number/animacy factor, the lexical semantics factor, and the frequency of mention factor.

A second model which uses an extended context is also implemented. It does not use the head environment (*h, t, l*) of ρ. Instead, it looks at the grammatical role of the candidates $\overrightarrow{G_w}$, the grammatical role of the pronoun G_ρ, and the sentence indices of candidates relative to the pronoun $\overrightarrow{S_w}$. Formally, this **syntactic-prominence** model is:

$$F(\rho)=\underset{a}{\operatorname{argmax}}P(A(\rho)=a \mid \rho,\vec{d}_H,\vec{W},\vec{M},S_\rho,f_\rho,G_\rho,\vec{G}_w,\vec{S}_w,ptn) \tag{6.3}$$

$$=\underset{a}{\operatorname{argmax}}P(\rho \mid a,W_a)P(d_a \mid a,f_\rho)P(G_{w_a} \mid a,G_\rho)P(S_{w_a} \mid a,f_\rho)\frac{P(M_a \mid a,G_\rho,S_\rho)}{P(M_a)}P(ptn \mid a) \tag{6.4}$$

This model captures, in addition to those in the basic model, the factors of grammatical salience and sentence recency. Unlike the basic model, it rids of lexical semantics.

The derivations of both (6.2) and (6.4) make use of a series of independence assumptions. In actuality very few things are truly independent. But in actual implementation, our assumptions are reasonable to make. In an effort to gain more gender/animacy knowledge, we find an unsupervised learning algorithm for automatically learning noun phrase gender information. We also have a smart program to learn near perfect animacy information by learning WHICH-class nouns and WHO-class nouns. A very simple version of transductive learning also proves to the helpful. After putting these pieces together, the syntactic-prominence model achieves 91.1% accuracy for anaphoric HE/SHE/IT pronouns and 87.5% for all anaphoric pronouns. In addition to achieving a competitive success rate, our system achieves complete autonomy.

6.2 Future research

It is well known that stress and intonation, which concern the information structure of sentences, can affect the coreference options of noun phrases in certain sentences. Examples like (6.1) and (6.2) show this effect:

(6.1) John hit ***Bill***. Then ***he*** was injured.
(6.2) John hit Bill. Then ĤE was injured.

When unstressed, the "*he*" in (6.1) refers to *Bill*. Coreference with "*Bill*" is blocked in (6.2) where "*he*" is stressed to express a contrastive meaning.

Stress and intonation are easily identifiable in spoken language. In writing, although currently (at least in the Wall Street Journal corpus) there is no marking to indicate this factor, it is entirely possible that these markings will be available in future collection of corpora. Either in the case of applying anaphora resolution to spoken language or in the case of handling stress marking in texts, this factor cannot be neglected. In our syntactic-prominence model (or the basic model), we would then need a parameter $Stress_p$, a Boolean value indicating whether or not the pronoun is stressed, and possibly a similar parameter for the candidate antecedents $\overline{Stress_w}$, a Boolean vector indicating if each candidate is stressed. A possible statistic to add on to equation (6.4) is $P(W_a \mid a, Stress_a, Stress_p)$.

While the mention counts approximate the topics of a discourse, they do not differentiate between global ones from local ones since mention counts are accumulated throughout the entire story. When processing a candidate antecedent, its mention counts indicate the number of times it has been referred to from the beginning. This count may be misleading particularly in the case where a local topic and a global topic interact with each other. Consider the following discourse:

6.3a ***Gerard Scannell***, the head of OSHA, said USX managers have known about many of the safety and health deficiencies at the plants for years, yet have failed to take necessary action to counteract the hazards.
b. A USX spokesman said the company had not yet received any documents from OSHA regarding the penalty or fine.
c. "Once we do, they will receive very serious evaluation," the spokesman said.
d. He said that, if and when safety problems were identified, they were corrected.
e. The USX citations represented the first sizable enforcement action taken by OSHA under ***Mr. Scannell***.
f. ***He*** has promised stiffer fines, though the size of penalties sought by OSHA have been rising in recent years even before he took office this year.

The pronoun we try to resolve is the first "*He*" in (6.3f). "*Mr. Scannell*" in (6.3e) has mention counts 2 at this point and the "*He*" in (6.3d) realizing the spokesman has been mentioned 3 times. The mention counts seem to indicate that *the spokesman* is the topic. In a sense this is correct but

it is only the topic of a local discourse segment consisting of sentences (6.3b – 6.3d). Sentence (6.3e) has a topic shift and reintroduces "*Mr. Scannell*" by a full expression. This is because although "*Mr. Scannell*" is the most salient entity in (6.3a), after the intervention of (6.3b – 6.3d), it has drifted into background. To bring it back into foreground, it needs to be referred to by a full expression. In fact, in (6.3e) the use of "*him*" instead of "*Mr. Scannell*" would have been misleading. If this ***re-introduction*** of entities and topic shift can be recognized, the mention counts can be used with more sophistication.

It is well known that pronouns require an antecedent which is highly salient in the context of utterance. This is the major motivation for the centering algorithm in which pronoun resolution in a given discourse context at least partly hinges on recognition of the center of attention at any given time. Salience can be achieved by different means. What the current approaches (all the algorithms we have discussed and many others in literature) do not fully address is the manner in which salience is utilized by the pronoun interpreter. Different manners result in different kinds of coherence. There is a well-known contrast between coherence by virtue of *narration* and coherence by *parallelism*, as illustrated by examples (6.4d) and (6.4d'):

6.4a. The three candidates had a debate today.
b. Bob Dole began by bashing Bill Clinton.
c. He criticized him on his opposition to tobacco.

d. Then Ross Perot reminded him that most Americans are also anti-tobacco.
d'. Then Ross Perot slammed him on his tax policies.

The preferred interpretation for "*him*" in (6.4d) is *Bob Dole* whereas in (6.4d') it is *Bill Clinton*. Without taking into account the meanings of the verbs "*remind*" and "*slam*" and the relationships between these verbs and those in the previous sentences, a consistent algorithm would choose the same antecedent for "him" in (4.6d) and (4.6d'). Among many other things, meanings of verbs seem to be especially important to anaphora resolution. For instance, it has been shown that for some verbs (Grober 1978) the subject is the cause of the action described by the verb, while for other verbs causality is attributed to the object. When two utterances are put together by a causal

connective like "*because*", a pronoun in the subject position of the second utterance will be coreferential with the argument of the first verb that is perceived as causal (Ehrlich 1980), as is the case for (6.5):

6.5 John admires ***Bill*** because ***he*** is reliable.

The pronoun "*he*" is coreferential with "*Bill*" because "*Bill*" is the cause of the "*admire*" action by **John**. Studies along this line of inquiry tend to focus on specific verbs. These studies have not attempted to provide a principled analysis of how knowledge of verbs affects pronoun interpretation. Research in verb semantics would eventually require a comprehensive theory of general knowledge.

One other common characteristic of all the approaches is that pronouns in a discourse are processed sequentially with no backtracking. One of the design goals of the centering algorithm is to model the preferences associated with a hearer/reader's immediate tendency to interpret pronouns. However occasions will and do arise where there is a need to backtrack to a correct interpretation. Consider the *Terry/Tony* example in (4.1) shown here in (6.6)

6.6a ***Terry*** really goofs sometimes.
b Yesterday was a beautiful day and ***he*** was excited about trying out his new sailboat.
c ***He*** wanted Tony to join ***him*** on a sailing expedition.
d ***He*** called him at 6am.
e He was sick and furious at being woken up so early.

It may be true that an addressee's first interpretation upon hearing/reading "*He*" in (6.6e) is to associate it to "*Terry*", but after processing the entire sentence, the inference from (6.6e) would force the interpretation to be changed to "*Tony*". Presently all algorithms process pronouns from the point of their occurrences and backwards. The need to process an entire sentence to recover the correct assignment is to be further investigated. Further processing of the rest of the sentence would invariably require some sort of content-based inferencing.

In conclusion, the results of this empirical study show that despite the success we have had in building an automatic anaphora resolution system, there are still many unanswered

questions in this area. The success of our system also indicates that there is a pressing need for the development of the pragmatic theory of anaphora[1] on the one hand, and the development of computational models based on these theories on the other. It is evident from the error analyses (section §3.4.4) that if we correct all the errors due to factors other than the lack of world knowledge, we could reduce the error rate by half and the accuracy could reach 95%. In order to get the remaining 5% we really need a theory and a model of world knowledge. This problem is at the heart of NLP and AI in general. The problems discussed in this section more and less require us to model world knowledge. But over the years, it has been impervious to numerous attempts. While keeping the problem in the back of our minds, we perhaps should move on to more tangible problems, problems that we could at least get our hands on. Instead of attacking the world knowledge directly, we could work on more concrete problems like learning verb semantics, learning word clusters, building machine translation systems, etc. All these subproblems will eventually lead to a better understanding of the world knowledge problem.

---

[1] Wasow (1986) has very nicely summed up this urgency:

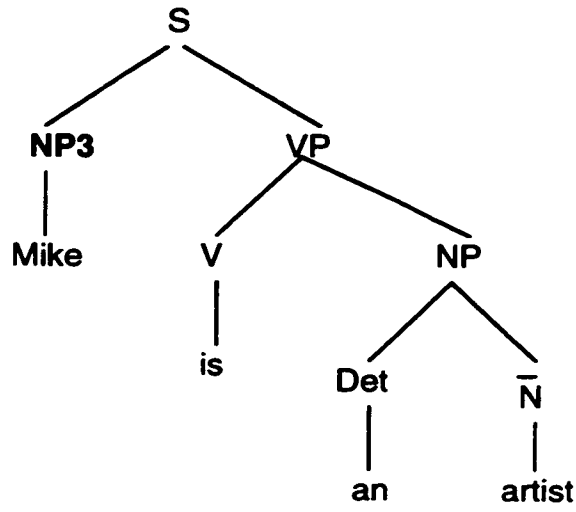
> Probably the most gaping hole in our present understanding of anaphora is the absence of any explicit theory of the pragmatic factors involved. There is, of course, a good deal of relevant literature, but there are no *theories* of the pragmatic aspects of anaphora which can compare in rigor or coverage with the available accounts of the syntactic and semantic factors.

Appendix A Hobbs' Algorithm

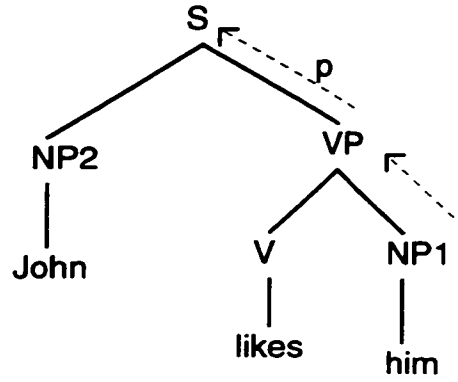
The input to the algorithm is a parse tree. The algorithm traverses the tree as follows.

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**, and call the path used to reach it *p*.
3. Traverse all branches below node **X** to the left of path *p* in a left-to-right, breadth-first fashion. Propose as the antecedent any NP node that is encountered which 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, the most recent first; each tree is traversed in a left-to-right, breadth-first manner, and when an NP is encountered, it is proposed as antecedent. If **X** is not the highest S node in the sentence, continue to 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 *p*.
6. If X is an NP node and if the path *p* to **X** did not pass through the $\overline{\text{N}}$ 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 node encountered. Propose any NP node encountered as the antecedent.
9. Go to step 4.

Consider the following example:

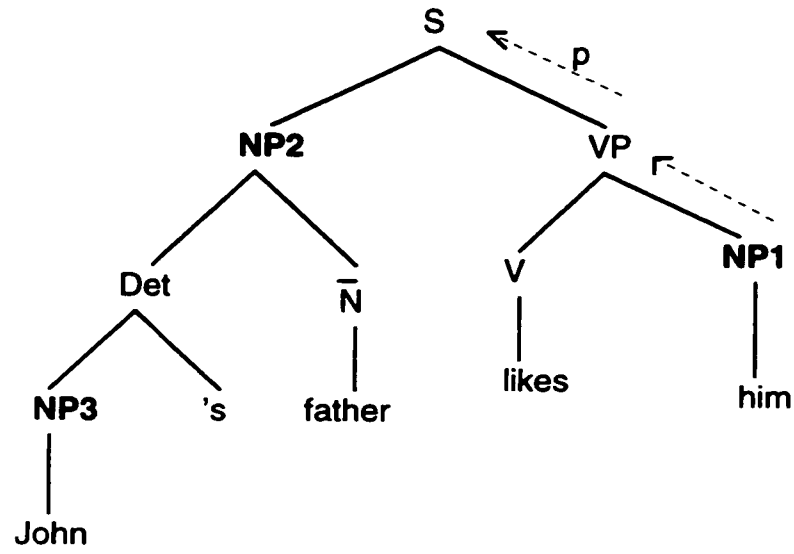

S
NP3
VP
Mike
V
NP
is
Det
N̄
an
artist

(A)


S
p
NP2
VP
John
V
NP1
likes
him

(B)

In resolving "him" in (B), we start at NP1 (step 1). We go up the tree to the first S node (step 2). A BFS search under the S node and to the left of the path p encounters NP2. But NP2 does not have an NP or S nod between it and the top S (step 3). Since S is the highest node in sentence (B), we go to the immediate previous sentence (A) (step 4). A BFS left-to-right search proposes NP3 (Mike) in (A) as an antecedent. In this case, it is the correct antecedent. Consider the parse tree in (C) which is a variation of (B):



(C)

Starting from NP1 (*him*), we go up the tree to the first S node encountered. We then search the tree to the left of the path p under the S node. NP2 (*John's father*) is ruled out because it is immediately dominated by the S node. The search continues and NP3 is encountered. NP3 is fine because it has NP2 between it and the S node. Hence NP3 (*John*) is proposed.

Appendix B

Distributions of antecedents' grammatical roles

$G_w \setminus G_p$	UMSBJ	ESBJ	NPSBJ	OBJ	PP	PPS	OTHER
UMSBJ	0.6714	0.4414	0.8	0.1967	0.3366	0.6	0.3873
ESBJ	0.1362	0.2920	0.1	0.2295	0.2970	0.2	0.3186
NPSBJ	0.1221	0.1150	0.1	0.1640	0.1535	0.2	0.1814
OBJ	0.0423	0.0759	0.1[1]	0.2951	0.0693	0.2[1]	0.0441
PP	0.0141	0.0736	0.1[1]	0.0984	0.1337	0.2[1]	0.0637
PPS	0.0141[1]	0.0023[1]	0.1[1]	0.0164	0.0099[1]	0.2[1]	0.0049[1]
OTHER	0.0141	0.0023	0.1[1]	0.0164[1]	0.0099	0.2[1]	0.0049

[1] The raw probability is actually 0. The number is smoothed to use the smallest probability in the column.

Appendix C Likelihood-ratio Test

The general formula of the Dunning statistic for the multinomial case is:

$$-2\log\lambda = 2\,[\log L(P_1, K_1) + \log L(P_2, K_2) - \log L(Q, K_1) - \log L(Q, K_2)]$$

where K stands for "count" and P stands for "probability" and

$$p_{ji} = \frac{k_{ji}}{\sum_j k_{ji}}$$

$$q_j = \frac{\sum_i k_{ji}}{\sum_{ij} k_{ji}}$$

$$\log L(P, K) = \sum_j k_j \log p_j$$

For our specific application, let

- ρ_j: the j^{th} pronoun class
- w_i: the i^{th} word
- n_ρ: total number of pronoun classes (= 7)
- n_w: total number of words
- $\neg w_i$: all words other than the i^{th} word

then

$$p_1 = \frac{K(\rho_i, w_i)}{K(w_i)}$$

$$p_2 = \frac{K(\rho_i, \neg w_i)}{K(\neg w_i)}$$

Likelihood-ratio Test (continued)

$$q = \frac{K(\rho_j)}{\sum_i^{n_w} K(w_i)}$$

$$\log L(P_1, K_1) = \sum_i^{n_w} \sum_j^{n_\rho} K(\rho_j, w_i) \log p_1$$

$$\log L(Q, K_1) = \sum_i^{n_w} \sum_j^{n_\rho} K(\rho_j, w_i) \log q$$

$$\log L(Q, K_2) = \sum_i^{n_w} \sum_j^{n_\rho} K(\rho_j, \neg w_i) \log q$$

Appendix D

Adjectives for Recognition of Pleonastics

A	amazing, apparent, appropriate, astonishing, awful, axiomatic
B	best, better, blasphemous
C	certain, charming, cheaper, clear, common, conceivable, costly, crucial
D	dangerous, decadent, demeaning, desirable, difficult, disingenuous, distasteful, doubtful
E	early, easier, easy, efficient, enjoyable, enough, entertaining, evident
F	fair, faster, favorable
G	good, great
H	hard, harder, healthy, horrible
I	imperative, important, impossible, inappropriate, inconceivable, incumbent, inevitable, inhumane, insulting, interesting, ironic, irresponsible
L	likely, logical
M	misguided, misleading
N	natural, necessary, nice, nonproductive
O	obvious, offputting, ok, okay
P	plain, painful, perfect, plausible, popular, possible, probable, proper, prudent, puzzling
R	rare, rational, realistic, reasonable, refreshing, ridiculous, right, risky
S	sad, safe, simple, sure, surprising
T	tempting, terrific, tough, tougher, trivial, true
U	unclear, uncommon, unfair, unfortunate, unlikely, unnecessary, unreasonable, unusual, unwise, useful
V	vital
W	weird, well-known, wiser, worrying, worse, worthy

Bibliography

Blaheta, D. and Charniak, E. (2000) *Assigning Function Tags to Parsed Text*, In *Proceedings of North America Association of Computational Linguistics*, Seattle, Washington

Brennan, S. E., Friedman, M. W., and Pollard, C. J. (1987) *A Centering Approach to Pronouns*, In *Proceedings of the 25th Annual Meeting of the Association of Computational Linguistics*, pp. 155 – 162, Stanford University, Stanford, California

Charniak, E. (1997) *Statistical Parsing with a Context-free Grammar and Word Statistics*, In *Proceedings of the 14th National Conference on Artificial Intelligence*, AAAI Press/MIT Press, Menlo Park, California

Charniak, E. (2000) *A Maximum-Entropy-Inspired Parser*, In *Proceedings of North America Association of Computational Linguistics*, Seattle, Washington

Chomsky, N. (1980) *On Binding*, In *Linguistic Inquiry*, Vol. 11, pp. 1 – 46, MIT Press, Cambridge, Massachusetts

Chomsky, N. (1981) *Lectures on Government and Binding*, Foris Publications, Dordrecht, Holland.

Cote, S. (1998) *Ranking Forward Looking Centers*, In *Centering Theory in Discourse*, Walker, A., Joshi, & Prince, E. (eds.) Oxford, Claredon Press

Dunning, T. (1993) *Accurate Methods for the Statistics of Surprise and Coincidence*, In *Association for Computational Linguistics*, pp. 61 – 74

Ehrlich, K. (1980) *Compreension of Pronouns*, Quarterly Journal of *Experimental Psychology*, Vol. 32, pp. 247 – 255

Ge, N., Hale, J., and Charniak, E. A *Statistical Approach to Anaphora Resolution*, In *Proceedings of ACL-98 Corpus-base NLP Workshop*, Montreal, Canada 1998

Gordon, P. C., Grosz, G. J., & Gilliom, L. A. (1993) *Pronouns, Names, and the Centering of Attention in Discourse*, In *Cognitive Science*, Vol. 17 pp. 311 – 347

Gordon, P.C., and Chan, D. (1995) *Pronouns, Passives and Discourse Coherence*, In *Journal of Memory and Language*, Vol. 34, pp. 216 – 231

Gordon, P. & Hendrick, R. (1997) Intuitive Knowledge of Linguistic Coreference, In Cognition, Vol. 62, pp. 325 - 370

Gordon, P. & Hendrick, R. (1998) The Representation and Processing of Coreference in Discourse, In Cognitive Science, Vol. 22, pp. 389 – 424

Grice, H. P. (1975) *Logic and Conversation*, In *Speech Acts, Syntax and Semantics*, Cole P. and Morgan J. L. (eds.) Vol. 3, pp. 41 – 58, Academic Press, New York

Grober, E.H., & Beardsley, Caramazza, A. (1978) Parallel Function Strategy in Pronoun Assignment, In Cognition, Vol. 6, pp. 117 – 133

Grosz, Barbara J. (1977) *The Representation and Use of Focus in Dialogue Understanding*. Technical Report 151, SRI International, Menlo Park, California

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983) *Providing a Unified Account of Definite Noun Phrases in Discourse*, In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, Cambridge, Massachusetts

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1995) *Centering: A Framework for Modelling the Local Coherence of Discourse*, In *Computational Linguistics*, Vol. 21 pp. 203 – 226

Grosz, B. J., & Sidner, C.L. (1986) *Attention, Intention, and the Structure of Discourse*, In *Computational Linguistics*, Vol. 12, pp. 175 – 204

Haegeman, L. (1991) *Anaphoric Relations and Overt NPs*, In *Introduction to Government and Binding Theory*. Oxford, Basil Blackwell

Hall, K. and Charniak, E. (2000) *A Probabilistic Model For Noun-Phrase Coreference*, Technical report, Department of Computer Science, Brown University

Hobbs, J. R. (1976) *Pronoun Resolution*, Research Report 76-1, City College, New York

Hobbs, J. R. (1979) Coherence and Coreference, In Cognitive Science, Vol. 3, pp. 67 – 90

Kamp, H., & Reyle, U. (1993) *From Discourse to Logic: Introduction to Model-Theoretic Semantics of Natural Language, Formal Logic, and Discourse Representation Theory.* Dordrecht: Kluwer Academic Publishers

Kehler, A. (1993) *Intrasentential Constraints on Intersentential Anaphora in Centering Theory*, presented at the Workshop on Centering Theory in Naturally Occurring Discourse, University of Pennsylvania, May 1993

Kuno, S. (1972) Pronominalization, Reflexivization, and Direct Discourse, in Linguistic Inquiry, Vol.3, pp. 161 – 195, MIT Press, Cambridge, Massachusetts

Kuno, S. (1972) Functional Sentence Perspective: A Case Study from Japanese and English, In Linguistic Inquiry, Vol. 3, pp. 269 – 320, MIT Press, Cambridge, Massachusetts

Lakoff, G (1968) *Pronouns and Reference*, In *Notes from the Linguistic Underground, Syntax and Semantics*, McCawley, J.D. (ed.) Vol. 13, pp. 275 – 345, Academic Press, New York

Langacker, R.W. (1969) *Pronominalization and the Chain of Command*, in D.A. Reibel and S. Schane(eds.), pp. 160-186

Lappin S. and Leass H. J. (1994) *An Algorithm for Pronominal Anaphora Resolution*, In *Computational Linguistics*, Vol. 20, pp. 535 – 561

Mitkov R., (1998) *Robust Pronoun Resolution with Limited Knowledge*, In Proceedings of 36th Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Vol. II, pp.869 – 875, Montreal, Quebec, Canada

Pollard, C., & Sag, I. (1992) Anaphors in English and the Scope of Binding Theory, In Linguistic Inquiry, Vol. 23, pp. 261 – 303

Reinhart, T. (1981) *Definite NP Anaphora and C-Command Domains*, In *Linguistic Inquiry*, Vol. 12, pp.605 – 635, MIT Press, Cambridge Massachusetts

Reinhart, T. (1983) *Anaphora and Semantic Interpretation*, Croom Helm Linguistics Series, Croom Helm Ltd. London & Canberra

Shank, R. (1973) Identification of Conceptualizations Underlying Natural Language, In Computer Models of Thought and Language, Shank R., & Colby, K. M. (eds.), San Francisco, California: Freeman

Strube, M. & Hahn, U. (1999) *Functional Centering – Grounding Referential Coherence in Information Structure*, In *Computational Linguistics*. Vol. 25, pp. 309 – 345

Turan, U. M. (1998) *Ranking Forward Looking Centers in Turkish: Universal and Language Specific Properties*, In *Centering Theory in Discourse*, Walker, A., Joshi, & Prince, E. (eds.) Oxford, Claredon Press

Walker M. A. (1989) *Evaluating Discourse Processing Algorithms*, In *Proceedings of the 27th Annual Meeting of the Association of Computational Linguistics*, pp. 251 – 261

Wasow, T. (1986) *Reflections on Anaphora*, In Lust (ed.) Vol. 1, pp. 107 – 122

Webber, B. L., (1979) *A Formal approach to Discourse Anaphora*, New York, Garland