Chapter 3
Early Approaches to Anaphora Resolution: Theoretically Inspired and Heuristic-Based

Massimo Poesio, Roland Stuckardt, Yannick Versley, and Renata Vieira

Abstract This chapter summarizes the most influential non-statistical approaches to anaphora resolution. Much of the very early work focused on *personal pronouns* and was based on theoretical proposals concerning anaphora and its interpretation developed in linguistics (e.g., the effect of syntax or semantics on anaphora) and/or psychology (e.g., on the effect of salience or commonsense knowledge). Such systems assumed the resolver would have *perfect information* available–e.g., on the syntactic structure of the sentence, or the properties of concepts and instances– and as a result, tended to be very brittle (a notable exception being Hobbs' 'naive' algorithm for pronoun resolution). In the first part of this Chapter we cover in detail some of these theoretically-motivated algorithms, such as Hobbs' and Sidner's, and briefly survey a number of other ones. The availability of the first corpora in the mid-90s (see Chapter 4) led to the development of the first systems able to operate on a larger scale, and to a widening of the range of anaphoric expressions handled. The fundamental property of these systems was the ability to carry out resolution on the basis of imperfect information only, using a variety of *heuristics*. In the second part of this Chapter, we cover a number of these heuristic-based algorithms. Some of the ideas developed in these heuristic-based systems have come back and are the basis for systems developed in the last few years; of these, we will discuss in some detail the Stanford Deterministic Coreference System.

Name of First Author
Name, Address of Institute, e-mail: name@email.address Massimo Poesio
University of Essex e-mail: poesio@essex.ac.uk

Roland Stuckardt
IT-Beratung • Sprachtechnologie • Medienanalyse, D-60433 Frankfurt am Main, Germany
e-mail: roland@stuckardt.de

Yannick Versley
Ruprecht-Karls-Universitt Heidelberg e-mail: versley@cl.uni-heidelberg.de

Renata Vieira
Universidade Católica do Rio Grande do Sul e-mail: renata.vieira@pucrs.br

3.1 Introduction

Between the '60s and the mid '90s a great number of computational models of anaphora resolution were developed, implementing the theories of the effect on anaphora of syntactic, commonsense, and discourse knowledge discussed in the Chapter 2. There are substantial differences between these models in terms of their theoretical assumptions (some models assume that anaphora resolution is entirely a matter of commonsense knowledge, others that it is almost entirely a matter of syntactic information) and their level of formality (some models are very linguistically and formally oriented, others are very much pragmatically oriented); but they covered quite a lot of ground, so that it is fair to say that most of what we know today about anaphora resolution was introduced as part of the development of these models. For this reason it makes sense to briefly cover these approaches before moving on more recent work. Of these proposals, this Chapter covers in some detail Hobbs' and Sidner's algorithms; and, more briefly, the commonsense-based algorithms of Charniak and Wilks, Lappin and Leass' algorithm, and other Centering-based algorithms. Our discussion will be short and focusing on the main ideas introduced in this work, many of which still valuable (and not yet incorporated in recent work). More in-depth discussion can be found in earlier surveys, such as [27, 55].

However, these models all have two aspects in common that set them apart from later work: (i) no large scale evaluation was attempted: the models were either purely theoretical, or the implementation was a proof of concept (the larger evaluation attempts, such as Hobbs', consider a few hundred cases); (ii) development was guided near-exclusively by the researcher's own intuitions, rather than by annotated texts from the targeted domain. The Message Understanding Conferences (MUC), and the development of the first medium-scale annotated resources, allowed researchers in the field to overcome these early limitations. Other key research, which marked as well the beginning of the resources-driven, or robustification phase, dealt with the issue of how to arrive at truly operational implementations of important anaphora resolution strategies—here, we will take an in-depth look at Stuckardt's ROSANA system that accomplishes robust syntactic disjoint reference. In the remaining part of this Chapter, we will then cover in some detail the two most influential heuristic pronoun resolution systems - Baldwin's CoGNIAC and Mitkov's MARS–and the Vieira and Poesio algorithm, one of the first to resolve *definite descriptions* on a large scale. We also review briefly the two best-performing systems that participated in the first 'coreference' resolution evaluation campaigns, FASTUS and LaSIE. Heuristic-based systems are still competitive; of the modern systems, we will discuss in some detail the Stanford Deterministic Coreference System.

Thus, we will survey the key approaches, algorithms, and systems of all past research stages as identified in Chapter 1, covering the knowledge-rich, domain-specific phase, the shallow processing phase, the consolidation phase, the resources-driven, or robustification phase, and the post-modern phase.

3.2 Hobbs' 'Naive' Syntax-Based Algorithm

We saw in Chapter 2 that (morpho) syntactic information plays an important role both in filtering certain types of interpretation (gender, binding constraints) and in determining preferred interpretations (subject assignment, parallelism). Several algorithms have been developed that incorporate these types of syntactic knowledge for anaphora resolution, in particular for the resolution of pronouns.

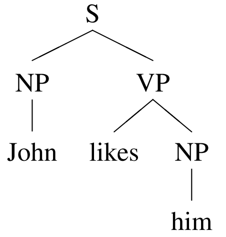
The earliest and best-known of these syntax-based algorithms is the pronoun resolution algorithm proposed by [28]. This algorithm, still often used as a baseline, traverses the **surface parse tree** breadth-first, left-to-right, and then going backwards one sentence at a time, looking for an antecedent matching the pronoun in gender and number. (See Figure 1.)

Algorithm 1 Hobbs' Algorithm

```
 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 node in the sentence then
 5:     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
 6: else
 7:     (X is not the highest node in the sentence) continue to step 9.
 8: end if
 9: 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.
10: if X is an NP node and if the path p to X did not pass through the N node that X immediately
    dominates then
11:     propose X as the antecedent
12: end if
13: 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.
14: if X is an S node then
15:     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.
16:     Propose any NP node encountered as the antecedent.
17: end if
18: Go to step 4
```

The algorithm incorporates both syntactic constraints, in particular from binding theory, and preferences, in particular subject and preference for first mentioned entities. Steps 2 and 3 ensure that no NP within the same binding domain as a pronoun will be chosen as antecedent for that pronoun, in that step 3 requires another NP or S node to occur in between the top node (node X) and any candidate: thus for example [NP John] will not be chosen as a candidate antecedent of pronoun *him* in example (3.1).

(3.1)



Because the search is breadth-first, left-to-right, NPs to the left and higher in the node will be preferred over NPs to the right and more deeply embedded, which is consistent both with the results of [19] concerning the effects of first mention and with the proposals and results of, e.g., [75] and [52] concerning the preference for antecedents in matrix clauses.

Hobbs was possibly the first anaphora resolution researcher to attempt a formal evaluation of his algorithm. He tested it (by hand, apparently) with 100 pronoun examples from three different genres (a historical text, a novel, and a news article) achieving 88.3% accuracy, on the assumption of perfect parsing. Hobbs also claimed that with the addition of selection restrictions, his algorithm could achieve 91.7% accuracy. Several subsequent larger-scale evaluations showed that when perfect syntactic knowledge is available (i.e., using syntactically hand-annotated corpora) the algorithm is very competitive, if not quite as accurate as these initial figures would suggest. [46] observed an accuracy of 82% over 360 pronouns from their corpus of computer manuals for their reimplementation of the algorithm. [76] found an accuracy of 76.8% over the 1694 pronouns in the Ge et al corpus of news text from the Penn Treebank, and of 80.1% over 511 pronouns from fictional texts. Hobbs' algorithm was also tested in a study by [51], who found reading time evidence for a left-to-right top-down breadth-first search for antecedents.

3.3 Approaches Based on Commonsense Knowledge

Although Hobbs developed his algorithm already in the 70s, it can in fact be seen as a very early indicator of a research paradigm shift, moving towards shallow processing strategies that are chiefly based on less knowledge-rich sources of evidence. But before moving further towards considering other increasingly knowledge-poor approaches, let's briefly look at some very early work, which accomplishes anaphora processing based on knowledge-rich strategies.

Much of the initial work on anaphora resolution both in computational linguistics and in psychology was devoted to providing an account of the effects of **commonsense knowledge** on anaphoric interpretation discussed in Chapter 2 and exemplified, e.g., by (3.2) (from [89]).

(3.2) a. The city council denied the women a permit because they feared violence.

b. The city council denied the women a permit because they advocated violence.

In this Section, we briefly discuss the most influential among these early models; for more recent work, see Chapter 14.

3.3.1 Charniak

In his dissertation [9], Charniak proposed a model of the use of inference in language comprehension largely motivated by problems in the interpretation of anaphora. For instance, the model aimed at explaining why in (3.3b) pronoun *it* is interpreted as referring to the piggy bank, whereas pronoun it in (3.3e) is interpreted as referring to the nickel.

(3.3) a. Janet wanted a nickel.
b. She went to her room to get her piggy bank, and found it.
c. At first, she couldn't get the nickel out.
d. So she shook the piggy bank again.
e. Finally, it came out.

Charniak developed a system called DSP (Deep Semantic Processing) taking as input the hand-coded assertions that a hypothetical semantic interpreter would produce for the sentences in an example like (3.3) and would carry out a number of deductive inferences that would resolve the anaphoric references as a byproduct. The deductive inferences were formulated in terms of PLANNER, one of the first languages for theorem proving developed in AI [25]. Charniak's account of the interpretation of (3.3e) involved a 'demon' (inference rule in PLANNER) that allowed to conclude that a possible binding for an object X that comes out is a coin Y contained in a piggy bank Z that gets shaken. An integral part of Charniak's proposal is an extensive theory of 'piggy banks' accounting for a number of other examples depending for their resolution on our knowledge about piggy banks.

Charniak's proposal was only partially implemented, not systematically evaluated, and has a number of known problems–e.g., the mechanism he proposed to choose among alternative inference paths in case more than one interpretation is possible is not very convincing– is possibly the first systematic attempt at providing a computational account of inference in anaphora resolution. In subsequent work (e.g., [10]) Charniak developed systems using frames in the sense of [53] to account for 'situational' bridging references such as *the aisles* and *the checkout counter* in (3.4).

(3.4) a. Jack went to the supermarket.
b. He got a cart and started going up and down the aisles.
c. He took the goods to the checkout counter and left.

3.3.2 Wilks

In [87, 88] and a number of other publications, Wilks presented an 'intelligent analyzer and understander of English' able to interpret English paragraphs and translate them into French via conversion into an interlingua based on Wilks' **Preference Semantics**. The system included an interesting treatment of anaphora resolution, and in particular of the role of semantics and commonsense knowledge.

Wilks's system did not use syntactic information. Instead, interpretation was carried out by slotting the **lexical templates** encoding the lexical semantics of the words in the sentence into **basic templates**–generic semantic templates for sentences (e.g., `[MAN FORCE MAN]`).[1] Selectional restrictions played an important role at this stage, e.g., in identifying the correct lexical template among those of an ambiguous word like *crook* to slot in the basic template for a sentence like *The policeman interrogated the crook*. At the next stage, **paraplates** expressing the interpretation of prepositions, other functional words, and connectives were used to combine together the basic templates thus enriched.

Wilks classified pronouns depending on the stage at which they are interpreted within this architecture. **Type A** pronouns are those that can resolved using 'shallow' semantic knowledge at the stage in which selectional restrictions are applied, as in (3.5), where the selectional restrictions specified by the lexical template for *hungry* are sufficient to choose *the monkeys* as the antecedent for *they*.

(3.5) Give the bananas to the monkeys, they must be very hungry.

Other cases, called **Type B** pronouns by Wilks, require deeper inference. For instance, in (3.6) , both whisky and glasses can 'feel warm'; in order to choose among antecedents it is necessary to carry out the inference that if object X gets drank by person Y, X ends in Y's stomach. According to Wilks, such inference can be carried out using lexical semantics only. To resolve such cases, Wilks' system entered a so-called **extended mode** in which such inferences were carried out by enriching the initial template through a process called **extraction**.

(3.6) John drank whisky from a glass. It felt warm in his stomach.

Finally, Wilks proposed that proper inference rules, that he called **Common Sense Inference Rules**, were needed to interpret pronouns like *it* in (3.7). The CSIR used in this case would conclude that when animal X ingests liquid Y it may be led to evaluate Y.

(3.7) John drank whisky from a glass. It was good.

Wilks' system was not properly evaluated, but the implementation of his approach to use semantics and commonsense knowledge for pronoun resolution was re-implemented and integrated with Sidner's account of salience [67] by Carter [8]; we will discuss this system below.

[1] In Preference Semantics, semantics is expressed in terms of a small number of **semantic primitives** like `FORCE`.

3.3.3 Hobbs, Hobbs and Kehler

The most systematic account of the use of inference in anaphora resolution can be found in Hobbs' work starting from the 'second algorithm' proposed in his dissertation [29][2] and expanded in a series of papers eventually leading to the theory of 'local pragmatics' [31, 32] incorporated in the 'interpretation as abduction' framework in which the whole of natural language interpretation was formulated as as abductive inference process [34].

To illustrate this proposal, let us consider the theory of the mutual effect of discourse connectives and coreference started with [30], and its treatment of example (3.2). We will follow the discussion in [42], a recent investigation of the predictions of Hobbs' theory using the methods of cognitive psychology. Let S1 be the sentence *The city council denied the demonstrators a permit*, and let this sentence have the semantic interpretation

(3.8) $deny(city\_council, demonstrators, permit)$

Let S2 be the continuation sentence (different in the two cases). According to Kehler *et al.* (and to Hobbs), connective *because* signals an **Explanation** relation between S1 and S2 in both versions of (3.2). Hobbs' formalization of **Explanation** is as follows:

> **Explanation**: Infer P from the assertion of S1 and Q from the assertion of S2, where normally $Q \rightarrow P$.

In order for the explanation of S1 in terms of S2 to be justified, some underlying axioms must exist that, simplifying a lot, could be expressed as the single following axiom:

(3.9) $fear(X,V) \wedge advocate(Y,V) \wedge enable\_to\_cause(Z,Y,V) \rightarrow deny(X,Y,Z)$

Axiom (3.9) says that if X (the city council) fears V (the violence), Y (the demonstrators) advocate V, and Z (the permit) enables Y to cause V, then we can 'plausiby infer' that X may deny Y to Z. According to Hobbs, in a situation in which axiom (3.8) has been asserted, and (3.9) is part of commonsense knowledge about the possible reasons for denial, abductive inference simultaneously establishes the existence of an **Explanation** while binding the council to X, the violence to V, the demonstrators to Y and therefore resolving pronoun *they* to the appropriate entity in both versions of S2. Clearly, such a theory does provide a convincing account for examples like (3.2), but a system based on such theory must be provided with axioms like (3.9).[3]

[2] In which the algorithm discussed in Section 3.2 and since known as "Hobbs' algorithm" was in fact presented as a baseline against which to evaluate the more sophisticated algorithm using commonsense knowledge.

[3] For an alternative account of the inference process leading to the establishment of coherence relations (although, to our knowledge, not of example (3.2)) see [4]. Systems making heavy use of such inferences for natural language interpretation were actually implemented by SRI, some of which also participated at the early MUC competitions, see e.g., [33, 2].

3.4 Salience: Sidner's Algorithm

The effect of recency on anaphoric interpretation is easy to notice; as a result, some mechanism to incorporate such preferences in anaphora resolution systems was present from the very early days and at least since the **history lists** of Winograd's SHRDLU [89]–data structures that store the potential antecedents most recently introduced first so that candidates are tested in the reverse order of introduction. As discussed in Chapter 2, however, simply choosing the most recently mentioned matching antecedent is not a particularly effective strategy, and already Hobbs' algorithm incorporates a more sophisticated notion of recency, sentence based and taking the first mention effect into account. The evidence about the effects on anaphoric interpretation of salience (as opposed to simple recency) discussed in Chapter 2, and in particular the work by [21], [49] and [66], motivated a great deal of research in computational linguistics producing models of anaphora resolution incorporating theories of salience [67, 65, 37, 6, 1, 84, 46, 75, 85, 86, 71, 72, 76]. Of these, the algorithms proposed in [67] and further developed by [8] and [75] arguably remain to this day the most detailed model of the effects of salience on anaphora resolution although their performance is unclear given that only a small-scale evaluation was attempted. We discuss these algorithms, which are representatives of the phase of (comparatively) shallow processing, in the present Section and more recent salience-based algorithms in the next.

3.4.1 Sidner's Computational Model of Focus

The central component of Sidner's theory is a discourse model with two key structural aspects:

- the organization of the entities in a semantic network inspired by the work of Charniak, although very few details about its organization are given in the original dissertation (see discussion of Carter's work in which this aspect of the theory was fleshed out below);
- data structures keeping track of which entities are currently most in focus. This aspect of the theory is the one which has had the greatest influence on subsequent research, in particular on the development of Centering (see next paragraph).

Sidner's theory of local focus is articulated around three main data structures: the **discourse focus**, her implementation of the notion of 'discourse topic' (see discussion in Chapter 2); the **actor focus**, accounting for the effects of thematic role preferences or subject assignment; and a ranked list of the entities mentioned in the last sentence. In addition, stacks of previous discourse foci, actor foci, and sentence foci lists are maintained. The first substantial part of Sidner's model are detailed algorithms that specify how each of these structures is updated as a discourse progresses.

Unfortunately there is no space here to discuss those algorithms (the algorithm for discourse focus update alone runs for two pages).

The second part of the model are algorithms that specify how the several focus structures she proposes are used in anaphoric interpretation. Sidner subscribed to an extreme version of the ’bottom up’ view of anaphora interpretation favored by psycholinguist: her model not only includes separate algorithms for each type of anaphoric expression, but also different algorithms for the same anaphoric expression depending on its (semantic) position. I.e., she doesn’t simply provide different algorithms for demonstrative and personal pronouns, but three different algorithms for personal pronouns in agent position, non-agent position, and possessive position. These algorithms differ regarding which local focus structures are accessed, and in which order. Again we do not have sufficient space for presenting all of these algorithms, but for illustration, the version of her algorithm for resolving 3rd person pronouns in non-agent position from [68] is shown in Fig. 2.

Algorithm 2 Sidner’s Algorithm for 3rd person Pronouns in Non-Agent Position

```
If the pronoun under interpretation appears in a thematic position other than AGENT, then
if there is no Discourse Focus (DF) then
    check if there are focus sets; if so, then hypothesize that the focus set serves as cospecification.
end if
if (Recency rule) the pronoun occurs first in the sentence and the last element of the Discourse
Focus List (DFL) is an NP then
    hypothesize a co specification between the pronoun and that DFL.
end if
if (Discourse Focus) the pronoun is plural and the DF is singular then
    hypothesize that the pronoun co specifies with the DF and an element of the DFL or the
    focus stack.
end if
Hypothesize that the pronoun co specifies with the DF.
if several objects associated with the DF are acceptable as co specification, and the pronoun
is plural then
    hypothesize a plural co specification; otherwise, predict that the pronoun was used ambiguously.
end if
if only one element associated with the DF is acceptable as co specification then
    hypothesize the co specification.
end if
Hypothesize DFL as co specification.
(Actor Focus) Hypothesize AF or PAF as co specification.
```

No evaluation of the theory was provided in Sidner’s thesis apart from discussing how it would work with several examples, but an evaluation was carried out by Carter.

3.4.2 Carter's SPAR system

Sidner's algorithms were partially implemented as part of the PAL system ('Personal Assistant Language Understanding Program') at MIT, and in the TDUS system at SRI (see [27] for an extensive discussion of PAL), but the most complete implementation of the theory was Carter's SPAR system [8].

SPAR is based on what Carter calls the **shallow processing hypothesis**, which limits calls to commonsense inference as much as possible since they are expensive and not very reliable. The system works by first producing all initial semantic interpretations of a sentence, expressed as formulas in Wilks' Preference Semantics formalism [87, 88] in which anaphors are left unresolved; and then attempting to resolve all the anaphors in each reading using Sidner's methods, and assigning to each reading a score which depends in part on how many anaphors have been successfully resolved and how many initial suggestions have been rejected. (It is at this point that Sidner's normal inference is invoked, rejecting interpretations for anaphors that do not satisfy some pretty basic commonsense knowledge see below.) The readings are then filtered, eliminating all those that do not satisfy configurational constraints (i.e., Reinhart's binding conditions); of those that remain, only the highest-scoring are accepted. If there is more than one such reading, then special inference mode is entered, in the form of Wilks' causal inference rules [88]. These rules are used to modify the previous scores. If still more than one reading has the same score, tie-breaking weak heuristics are used. In what follows, we will briefly discuss Carter's modifications to Sidner's theory, how SPAR integrates salience and commonsense knowledge, and the results of his evaluation.

Carter's modifications to Sidner's theory

The first modification to Sidner's theory proposed by Carter is to eliminate the Recency Rule (see algorithm 2) which, according to him, systematically led to worse results. The second modification concerns the treatment of intrasentential anaphors, for which Sidner made no provision. Other researchers who tackled this problem—in particular Suri and McCoy (discussed next) and Kameyama—proposed to deal with intra-sentential anaphora by updating the focus registers at additional points inside the sentence, instead of just at the end of each sentence. By contrast, in SPAR intrasentential anaphors are handled by making some intrasentential antecedents temporarily available by adding them to the DFL and AFL, and by modifying the rules for resolving third-person pronouns so that they also consider these antecedents, in addition to those stored in the other focus registers proposed by Sidner. One advantage of this approach is that it can also be used for intraclausal anaphora.

Interaction with Commonsense Reasoning

Carter's approach to using reasoning to resolve pronouns follows from Wilks', who proposed that the following steps are followed:

1. collect the candidates that match the pronoun syntactically;
2. apply selectional restrictions;
3. use analytic inference rules to derive equivalent propositions and then try to derive an interpretation for pronouns by matching these propositions with the original ones;
4. use commonsense inference rules again to infer new propositions and try to find matching antecedents.

These steps are in a progression from strong syntactic constraints to weak commonsense inferences. Also, the first two steps can be performed separately on the candidates for each pronoun, whereas the last two can only be performed starting from a complete interpretation for the sentence (i.e., one in which an hypothesis about each anaphor has been made). As a consequence, Carter proposes to identify what Sidner calls normal inference mode with the first two steps, which are then performed for every pronoun; and what she calls special inference mode with the second two steps, which are only performed after a set of candidates for all anaphoric expressions has been constructed, and if more than one interpretation is still possible.

Evaluation

SPAR was tested with two types of texts. The first set includes 40 short texts (one to three sentences), written by Carter himself to test SPAR's capabilities; all anaphors in these texts are resolved correctly. The second set consists of 23 texts written by others, of average length nine sentences, and containing 242 pronouns in total; of these, 226 (= 93%) are resolved correctly.

3.4.3 Suri and McCoy

Suri and McCoy [75] proposed a revision of Sidner's theory called RAFT/RAPR. Just as in Sidner's theory, two foci are maintained for each sentence in RAFT/RAPR: the Subject Focus (SF) (corresponding to Sidner's Actor Focus) and the Current Focus (CF) (corresponding to Sidner's Discourse Focus). The two foci often refer to distinct objects, although that need not be the case.

Another characteristic that RAFT/RAPR inherits from Sidner's theory is that in addition to a Current Focus, a Subject Focus, and two lists of Potential Foci, the data structures assumed by the pronoun resolution and focus tracking algorithms also include stacks of all the information computed in previous sentences, i.e., a CF stack, a SF stack, a PFL stack, and a PSFL stack. Finally, the pronoun resolution algorithm proposed by Suri and McCoy, like Sidner's, is based on the assumption that hypotheses are generated one at a time, and accepted or rejected by commonsense reasoning.

The first change to Sidner's theory introduced by Suri and McCoy is the replacement of thematic relations with grammatical functions both in the Focusing Algorithm and in the Pronoun Interpretation Algorithm. Thus, the SF is defined as the subject of the sentence; the FA for computing the CF relies on syntactic notions rather than thematic roles. And in Suri and McCoy's version of the PIA, unlike in Sidner's, a distinction is made between subject and non-subject pronouns, rather than between AGENT and non-AGENT ones. A second important modification is that Suri and McCoy, like Carter, extend Sidner's algorithms to include complex sentences.

3.5 Other Salience-Based Algorithms: Centering based and Activation-based Models

Two main families of computational models of salience alternative to Sidner's have been developed in Computational Linguistics. Most of the best known work has been developed within the framework of Centering theory [22], which has also been the theoretical foundation for a great deal of work in natural language generation [12, 45, 41]. As discussed in Chapter 2, Centering was originally intended as a simplification of Sidner's model in which only one focus was present, although in practice a 'second focus' is still present in most algorithms based on Centering. In anaphora resolution, the two best known algorithms based on Centering theory were developed by Brennan *et al.* [6] and by Strube and Hahn [72].

The second family includes models which view salience as a graded notion: instead of as discrete set of 'foci', such models assign a degree of salience to all discourse entities. The earliest such model known to us is from Kantor [27] but the best-known algorithm of this type is RAP by Lappin and Leass [46], which we will discuss in some detail in this Section, whereas in Section 3.6 we will discuss in detail the ROSANA algorithm that aims to make Lappin and Leass' approach work in knowledge-poor settings. Referring to the terminology introduced in Chapter 1, we are now considering representatives of the consolidation phase and the resources-driven, or robustification phase.

3.5.1 The Centering algorithm by Brennan, Friedman, and Pollard

The algorithm proposed by Brennan *et al.* (henceforth: BFP) takes as input utterance u_n and updates the local focus by choosing the pair

$$\langle CB_n, [CF_n^1, \ldots, CF_n^m] \rangle$$

which is most consistent with the claims of Centering. This is done in a generate-filter-rank fashion:

1. Produce all possible $\langle CB_n, [CF_n^1, \ldots, CF_n^m] \rangle$ pairs. This is done by computing the CFs–which in turn involves generating all interpretations for the anaphoric expressions in utterance u_n–and ranking them.
2. Filter all pairs which are ruled out either by hard constraints (e.g., of the binding theory) or by the constraints of Centering (see Chapter 2): that if any CF is pronominalized, the CB is; and that the CB should be the most highly ranked element of the CF list of u_{n-1} that is realized in u_n. The CFs are ranked according to grammatical function, with subjects ranking more highly than objects, and these than adjuncts.
3. Finally, the remaining pairs are ranked according to the preferences among transitions: namely, that maintaining the same CB as the most highly ranked (**continuing**) is preferred over maintaing the CB, but in less prominent position (**retaining**) which in turn is preferred over changing the CB (**shifting**).

The BFP algorithm has been extremely influential. Some of its features are grounded in solid empirical evidence–e.g., [60] found very few exceptions for the preference for pronominalizing the CB if any other entity is pronominalized–but other characteristics found less empirical verification: e.g., there is little behavioral evidence for the preferences among transitions [19] and real texts do not appear to be consistent with such preference either [60]. BFP did not themselves provide an evaluation of the algorithm, but [84] evaluated it by hand comparing its performance for pronouns with that of Hobbs' algorithm, over the same texts used by Hobbs. The BFP algorithm performed slightly better than Hobbs' on the narrative texts (90% accuracy vs. 88%), whereas Hobbs' algorithm performed slightly better over the task-oriented dialogues (51% vs. 49%) and clearly better with the news data (89% vs. 79%), the difference coming from Hobbs' algorithm preference for intrasentential antecedents, whereas the BFP algorithm tended to prefer intersentential ones. However, Tetreault's more extensive (and automatic) evaluation in [76] suggests that the performance of Hobbs' algorithm is actually rather better than that of the BFP algorithm: Hobbs achieved 80.1% accuracy with fictional texts vs. 46.4% for BFP, whereas with news articles, Hobbs achieved 76.8% accuracy vs. 59.4% for BFP.

In the algorithm proposed by [72], ranking by grammatical function is replaced by 'functional' ranking, i.e., ranking according to the taxonomy of given-new information proposed by [61]: (hearer) old entities (i.e., anaphoric entities and entities referred to using proper names) are ranked more highly than 'mediated' (i.e., bridging) references, and these more highly than hearer-new entities. Strube and Hahn evaluated the performance of their algorithm by hand for both English and German, using both narrative and newspaper texts for a total of arond 600 pronouns for each language, and comparing the accuracy with that of the BFP algorithm. The performance using functional ranking was higher than using grammatical function ranking for both languages. For English, they obtained 80.9% accuracy as opposed to 76% for BFP, whereas for German, they achieved 83.7% with functional ranking vs. 74.8% with grammatical function ranking. The good performance of functional ranking was confirmed by the corpus study of [60], which found that the parameter configuration with functional ranking was the one for which most of Centering's hypotheses were supported by the evidence.

3.5.2 The Graded Salience approach of Lappin and Leass

An alternative account of salience effects is centered around the notion of **activation**. Whereas Sidner's focusing theory and Centering account for salience effects by stipulating a discrete number of items in focus (the discourse focus, the CB, etc), activation-based models assume that every discourse entity has a certain level of activation on a graded scale (often values in the range 0...1), updated after every utterance, and that it is this level of activation that determines the likelihood of that entity being referred to. Activation-based models are less discussed, but in fact most commonly used in anaphora resolution systems than discrete models of salience.

The first known system of this type was proposed by [40] (see also [27] for discussion), but the best known models are the MEMORY system proposed by [1] (which also includes a detailed theory of semantic network use in anaphora resolution), and the RAP pronoun resolution algorithm proposed by [46], that builds on Alshawi's work but includes several innovations, above all the first extensive treatment of expletives, and has become one of the best known pronoun resolution algorithms in CL. RAP also incorporates a sophisticated treatment of binding constraints.

Lappin and Leass's algorithm is another good example of the *generate-filter-rank* model of anaphora resolution. RAP takes as input the output of a full parser, and uses the syntactic information to filter antecedents according to binding constraints, specifically (i) antecedents of non-reflexives when the pronoun occurs in the argument, adjunct or NP domain of the potential antecedent (e.g. *John$_i$ wants to see him$_{*i}$*, *She$_i$ sat near her$_{*i}$*, *John$_i$'s portrait of him$_{*i}$*), and (ii) non-pronominal antecedents that are contained in the governing phrase of the pronoun (*He$_i$ believes that the man$_{*i}$ is amusing*, *His$_i$ portrait of John$_{*i}$*). Reflexive pronouns are instead resolved to an antecedent that fulfills the binding criteria.

Of all the candidates that pass the syntactic filter and are number and gender compatible with the pronoun, the one with the highest *salience weight* is selected, breaking ties by selecting the closest antecedent.

Each mention receives an initial salience weight, consisting of:

- A *sentence recency* weight, which is always 100.
- Additional weights for mentions not occurring in dispreferred position such as embedded in a PP (*head noun emphasis*, 80), or in a topicalized adverbial PP (*Non-adverbial emphasis*, 50).
- A weight depending on the grammatical function (80 for subjects, 50 for direct objects, 40 for indirect objects or oblique complements). Predicates in existential constructions also receive a weight (70).

The weight for each antecedent mention is halved for each sentence boundary that is between anaphor and then summed across all the members of the coreference chain of a candidate. To this salience value for the discourse entity, two local factors are added: one for parallelism of grammatical roles (35) and a penalty for cataphora (-175), which is applied to antecedent candidates that appear *after* the anaphoric pronoun.

Lappin and Leass evaluated RAP using 360 previously unseen examples from computer manuals. RAP finds the correct antecedent for 310 pronouns, 86% of the total (74% of intersentential cases and 89% of intrasentential cases). Without the combination of salience degradation and grammatical function/parallelism preferences, the performance gets significantly worse (59% and 64%, respectively), whereas other factors seem to have a much smaller impact (4% loss in accuracy for a deactivation of the coreference chains features, 2% loss for a deactivation of the cataphora penalty). By contrast, their reimplementation of Hobbs's algorithm achieves 82% accuracy on the same data.

3.5.3 The shallow implementation of RAP by Kennedy and Boguraev

Lappin and Leass use deep linguistic information in three places: firstly, to determine binding-based incompatibility and restrictions on the resolution of reflexives; secondly, to assign salience weights based on grammatical functions; thirdly, they use the parser's lexicon to assign the gender of full noun phrases. An approach based on shallow processing would have to approximate the syntax-based constraints based on the information in partial parses, and use a heuristic approach to reach full coverage for gender determination. [43] use a Constraint Grammar parser that determines morphological tags and grammatical functions and allows the identification of NP chunks, but does not yield enough information for constructing a complete tree, and report 75% resolution accuracy for news text, citing incomplete gender information and quoted passages as the most important source of errors.

Kennedy and Boguraev don't provide formal descriptions of the rules they employ for robustly emulating the syntactic disjoint reference conditions on the Constraint Grammar parses. However, as this is definitly a key issue for robust, truly operational anaphora resolution, we will take a look on another thorough solution below in section 3.6, providing an in-depth description of the ROSANA algorithm by Stuckardt, which works on potentially fragmentary *full* parses.

3.5.4 Centering ff.: the algorithms by Strube and Tetreault

The algorithms proposed by [71] and [76] were inspired by Centering, but are in fact a version of the activation models in which activation scores (a partial order) are replaced by a list (a total order).

Tetreault's algorithm, Left-to-Right Centering (LRC), shown in 3, is the simplest and yet arguably the most effective algorithm inspired by Centering. It combines the idea of ranking of CFs from Centering with several ideas from Hobbs' algorithm.

Algorithm 3 Tetreault's LRC Algorithm

```
for all $U_n$ do
  parse $U_n$
  for all $CF_i$ in the parse tree of $U_n$ traversed breadth-first, left-to-right do
    if $CF_i$ is a pronoun then
      search intrasententially in CF-partial($U_n$), the list of CFs found so far in $U_n$, an antecedent that meets feature and binding constraints.
      if found matching antecedent then
        move to the next pronoun in $U_n$
      else
        search intersententially in CF($U_{n-1}$) an antecedent that meets feature and binding constraints.
      end if
    else
      add $CF_i$ to CF-partial($U_n$)
    end if
  end for
end for
```

Tetreault evaluated his algorithm using a corpus of texts from two genres: news articles (a subset of the Penn Treebank containing 1694 pronouns annotated by [18]), and fictional texts (also from the Penn Treebank, for a total of 511 pronouns). Tetreault also compared his algorithm with a variety of baselines, and with reimplementations of the BFP and Hobbs algorithms. On news articles, LRC achieved an accuracy of 80.4%, as opposed to 59.4% for BFP and 76.8% for Hobbs. On fiction, LRC achieved 81.1% accuracy, compared with 80.1% of Hobbs and 46.4% of BFP.

3.6 Robust syntactic disjoint reference: Stuckardt's ROSANA system

In recognizing that the Lappin & Leass algorithm [46] is not applicable in knowledge-poor scenarios as it requires full and unambiguous parses, the ROSANA[4] algorithm by Stuckardt aims at generalizing the *generate-filter-rank* approach in order to make

[4] ROSANA = **Ro**bust **S**yntax-Based Interpretation of **Ana**phoric Expressions

it work on partial (in the sense of fragmentary) parses. The focus is on respectively restating the syntactic disjoint reference conditions (derived from principles A, B, C and the i-within-i constraint of Binding Theory (BT)) so that as much configurational evidence as possible is exploited. Compared to the above-mentioned approach of Kennedy & Boguraev [44], which employs heuristic rules to partially reconstruct constituent structure from the results of a shallower preprocessing, it is thus aimed at exploiting syntactic evidence in the best possible way. ROSANA resorts to the potentially fragmentary parses derived by the robust FDG parser for English of Järvinen & Tapanainen (1997: [36]).[5]

In Figure 3.1, the *filtering* and *ranking/selection* phases of the ROSANA algorithm are specified. There are three main steps: *1. candidate filtering*, *2. candidate scoring and sorting*, and *3. antecedent selection*. In the filtering step, standard restrictions such as number-gender agreement and syntactic disjoint reference criteria are applied. In the scoring and sorting (= ranking) step, a numerical plausibility score comprising various factors is computed for each remaining candidate; in particular, this includes a graded salience weight similar to that employed by Lappin & Leass. Finally, in the selection step, for each anaphor, the highest scoring candidate that has survived filtering is chosen; as there might be interdependencies between the individual antecedent decisions, special care is taken to avoid conflicting antecedent assignments.

It would be beyond the scope of this exposition to describe these steps in full detail; the reader is referred to Stuckardt (2001: [73]) for further information on, e.g., how ROSANA implements graded salience, and how it identifies anaphors to be resolved and antecedent candidates in the preceding *generate* phase. However, some more space shall be allocated to discussing the key issue of robust syntactic disjoint reference implementation. In the respective filtering step 1b, which, by definition, considers intrasentential candidates only, two cases are distinguished: anaphor and candidate occur in the same subtree of the (possibly fragmentary) parse, vs. anaphor and candidate occur in different subtrees. It is the latter condition that signifies the application case of a set of **rule patterns** specifically designed to emulate the syntactic disjoint reference conditions on incomplete parses, i.e., parse fragments as typically occurring due to structural (PP, adverbial clause, etc.) ambiguities. To look at one particular case, *rule pattern [F2]*[6]

$$* \qquad \{\ldots F_i = [\ldots bn(\gamma)(\ldots \gamma_{typeA/B/C} \ldots)..],..,F_j = [\ldots bc(\alpha)(\ldots \alpha_{typeA} \ldots)..]\ldots\}$$

applies for reflexive (= BT type A) pronouns α that occur in syntactic fragments F_j which contain their binding categories $bc(\alpha)$. Any candidate γ of arbitrary BT type (A, B, or C) that occurs in a different fragment F_i containing its branching node $bn(\gamma)$ can be discarded (pattern prediction: $*$) since it is impossible to structurally conjoin the two fragments in a way that γ, as required by BP A of α, locally binds α: in case the anaphor's fragment is subordinated under the candidate's fragment,

[5] The FDG parser is the predecessor of the commercially available Connexor Machinese Syntax parser (`www.connexor.com`).

[6] Notational conventions: round brackets delimit constituents; square brackets emphasize fragment (= parse subtree) boundaries.

1. *Candidate Filtering*: for each anaphoric NP α, determine the set of admissible antecedents γ:

   a. verify morphosyntactic or lexical agreement with γ;

   b. if the antecedent candidate γ is intrasentential:
   - if α and γ belong to the same syntactic fragment, then verify that
     - i. the binding restriction of α is constructively satisfied,
     - ii. the binding restriction of γ is not violated,
     - iii. no i-within-i configuration results;
   - else (α and γ belong to different syntactic fragments) *try the rule patterns*:
     - iv. if one of the patterns [E2], [E3a], [E3b], [E4], or [F2] is matched, then some binding restrictions are violated,
     - v. else if one of the two i-within-i rule patterns applies, then some binding restrictions are violated,
     - vi. else if pattern [E1a], [E1b], or [F1] applies, then the binding restrictions of α and γ are satisfied,
     - vii. else (*no rule pattern applies*) assume heuristically that the binding restrictions of α and γ are satisfied;

   c. if α is a type B pronoun, antecedent candidate γ is intrasentential, and, with respect to surface order, γ *follows* α, verify that γ is *definite*.

2. *Candidate scoring and sorting*:

   a. for each remaining anaphor-candidate pair (α_i, γ_j): based on a set of preference heuristics, determine the numerical plausibility score $v(\alpha_i, \gamma_j)$.
   If the binding-theoretic admissibility was approved *heuristically* in step 1(b)vii, then reduce the plausibility score $v(\alpha_i, \gamma_j)$ by a constant value;

   b. for each anaphor α: sort candidates γ_j according to decreasing plausibility $v(\alpha, \gamma_j)$;

   c. Sort the anaphors α according to decreasing plausibility of their respective best antecedent candidates.

3. *Antecedent Selection*: consider anaphors α in the order determined in step 2c. Suggest antecedent candidates $\gamma_j(\alpha)$ in the order determined in step 2b.
   Select $\gamma_j(\alpha)$ as candidate if there is no interdependency, i.e. if

   a. the morphosyntactic features of α and $\gamma_j(\alpha)$ are still compatible,

   b. for all occurrences $\delta_{\gamma_j(\alpha)}$ and δ_α the coindexing of which with $\gamma_j(\alpha)$ and (respectively) α has been determined in the *current* invocation of the algorithm: the coindexing of $\delta_{\gamma_j(\alpha)}$ and δ_α, which results transitively when chosing $\gamma_j(\alpha)$ as antecedent for α, does neither violate the binding principles nor the i-within-i condition, i.e.
   - if $\delta_{\gamma_j(\alpha)}$ and δ_α belong to the same syntactic fragment, then, for both occurrences, verify the respective binding conditions and the i-within-i condition according to steps 1(b)ii and 1(b)iii,
   - else if $\delta_{\gamma_j(\alpha)}$ and δ_α belong to different syntactic fragments, then proceed according to steps 1(b)iv, 1(b)v, 1(b)vi, and 1(b)vii (with the exception of the rule patterns [F2], [E2], and [E4], by means of which binding principle A is *constructively* verified).

   (The case $\delta_{\gamma_j(\alpha)} = \gamma_j(\alpha) \wedge \delta_\alpha = \alpha$ does not need to be reconsidered.)

Fig. 3.1 Stuckardt's ROSANA algorithm — candidate filtering and ranking/selection phases

[F1] $\surd$ $\{\ldots F_i = [\ldots bc(\gamma)(\ldots \gamma_{typeB} \ldots) \ldots], .., F_j = [\ldots bc(\alpha)(\ldots \alpha_{typeB} \ldots) \ldots] \ldots\}$
[F2] $*$ $\{\ldots F_i = [\ldots bn(\gamma)(\ldots \gamma_{typeA/B/C} \ldots) ..], .., F_j = [\ldots bc(\alpha)(\ldots \alpha_{typeA} \ldots) ..] \ldots\}$

[E1a] $\surd$ $\{\ldots F_d = [\ldots \gamma_{typeA/B/C} \ldots], \ldots, F_e = [\ldots bc(\alpha)(\ldots \alpha_{typeB} \ldots) \ldots] \ldots\}$
[E1b] $\surd$ $\{\ldots F_d = [\ldots \alpha_{typeB/C} \ldots], \ldots, F_e = [\ldots bc(\gamma)(\ldots \gamma_{typeB} \ldots) \ldots] \ldots\}$
[E2] $*$ $\{\ldots F_d = [\ldots \gamma_{typeA/B/C} \ldots], \ldots, F_e = [\ldots bc(\alpha)(\ldots \alpha_{typeA} \ldots) \ldots] \ldots\}$
[E3a] $*$ $\{\ldots F_d = [\ldots \gamma_{typeA/B/C} \ldots], \ldots, F_e = [\ldots \alpha_{typeC} \ldots] \ldots\}$,
if γ c-commands α regardless of the attachment choice
[E3b] $*$ $\{\ldots F_d = [\ldots \alpha_{typeA/B/C} \ldots], \ldots, F_e = [\ldots \gamma_{typeC} \ldots] \ldots\}$,
if α c-commands γ regardless of the attachment choice
[E4] $*$ $\{\ldots F_d = [\ldots \alpha_{typeA} \ldots], \ldots, F_e = [\ldots bn(\gamma)(\ldots \gamma_{typeA/B/C} \ldots) \ldots]$

Fig. 3.2 rule patterns employed by ROSANA for robust binding constraint verification

[F1]	BP B of α / γ is satisfied	γ does not *locally* bind $\alpha \wedge \alpha$ does not *locally* bind γ
[F2]	BP A of α is violated	γ does not *locally* bind $\alpha \vee \gamma$ does not c-command α
[E1a]	BP B of α is satisfied	γ does not *locally* bind α
[E1b]	BP B of γ is satisfied	α does not *locally* bind γ
[E2]	BP A of α is violated	γ does not *locally* bind α
[E3a]	BP C of α is violated	γ c-commands α
[E3b]	BP C of γ is violated	α c-commands γ
[E4]	BP A of α is violated	γ does not c-command α

Fig. 3.3 binding-theoretic background of the ROSANA rule patterns

the presence of $bc(\alpha)$ ensures that no relation of *local* binding holds; in the opposite case, the presence of $bn(\gamma)$ rules out that a relation of c-command may be established.

The complete set of patterns employed by ROSANA to robustly implement the syntactic disjoint reference conditions is displayed in figure 3.2;[7] their binding-theoretic background is explicated in figure 3.3. [8]

In the antecedent selection step, individual antecedents are iteratively chosen in the order of decreasing plausibility, employing a greedy strategy. Two additional tests check for compatibility with the decisions made so far. In particular, step 3b accounts for the proper verification of the syntactic disjoint reference conditions,

[7] Between fragments named F_d and F_e, an embedding relation is assumed, requiring that the parser provides the additional information that the latter fragment is subordinated to the former.

[8] The two additional basic patterns that are employed in step 1(b)v for verifying the i-within-i condition of BT are specified in Stuckardt (2001: [73])

which characterize valid index *distributions* rather than valid individual relations of anaphoric resumption. To give an example, in the case

John informs Jerome that he will call him tomorrow.

the antecedent decisions *John* ← *he* and *John* ← *him* are both individually admissible, as the binding conditions of anaphor (BP B) and antecedent (BP C) are satisfied; however, combining these decisions would lead to the unacceptable index assignment

* *John*$_i$ *informs Jerome*$_j$ *that he*$_i$ *will call him*$_i$ *tomorrow.*

as the binding condition of the type B pronoun *him* gets *transitively* violated. Again, the test distinguishes between whether anaphor and antecedent candidate occur in the same or in different parse fragments, applying the above patterns where appropriate. Regarding the binding condition of type A pronouns (reflexives, reciprocals), care has to be taken not to be overly restrictive, taking into account that further non-local coindexings are admissible as long as there is one local antecedent as constructively demanded by BP A:

John$_i$ *says that he*$_i$ *shaves himself*$_i$.

ROSANA has been fully implemented[9] and automatically evaluated on a mid-sized corpus of referentially annotated news agency press releases. Evaluation has been carried out employing diverse measures, including model-theoretic coreference scoring ([83], $(P,R) = (0.81, 0.68)$), immediate antecedents (accuracy of 0.71 for third-person non-possessives, and 0.76 for third-person possessives), and non-pronominal antecedents (accuracy of 0.68 and 0.66, respectively). According to an error case breakdown by Stuckardt, none of the 7 incorrect antecedent choices that are due to failures of the syntactic disjoint reference strategy (out of a total of 246 wrong antecedent choices) are caused by wrong predictions of its robust operationalization, which is still partly heuristic; rather, these failures are identified to be caused by wrong (in contrast to partial) parsing results, among which cases of wrongly interpreted ambiguous relative clauses are prevailing. Stuckardt thus concludes that the robust implementation of syntactic disjoint reference is nearly optimal, identifying the possibility of a further slight improvement based on an employment of a more defensive parsing strategy.

3.7 Heuristic Approaches: Pronoun Resolution

Both ROSANA and Boguraev and Kennedy's reimplementation of RAP are early representatives of the resources-driven, or robustification phase, which began around

[9] See `www.stuckardt.de/index.php/anaphernresolution.html` for details about the distribution; there is as well an implementation available for the German language, which works on the output of the Connexor Machinese Syntax parser.

1995. In those years, the focus of Computational Linguistics started to shift towards algorithms and systems whose performance could be evaluated over larger datasets. In anaphora resolution, as well, the ability to carry out larger-scale evaluation started to be considered essential. This led to the development of a new generation of algorithms and systems that could be evaluated in this way. Such algorithms typically did not assume that perfect syntactic knowledge or commonsense knowledge were available, as neither large-scale full parsing, nor large-scale lexical resources, were possible at the time. Instead, **heuristic** methods were employed to get around these limitations.

In this Section we discuss two other well known heuristic algorithms for pronoun resolution, which are typical examples for the resources-driven, or robustification phase: CogNIAC, due to Breck Baldwin [5] and MARS, due to Mitkov [54]. In the next, we will discuss the Vieira/Poesio algorithm for definite description resolution, which is also one of the first examples of an approach based on machine learning.

3.7.1 CogNIAC

CogNIAC was designed around the assumption that Hobbs' conclusion in [29] that anaphora resolution necessarily requires commonsense knowledge was incorrect or, at least, overly pessimistic, and that there is a sub-class of pronominal anaphora that does not require general purpose reasoning. Like most other **knowledge poor** systems discussed above and in this Section, CogNIAC only requires part-of- speech tagging, recognition of noun phrases, and agreement information; it can use full parse trees if available.

What makes CogNIAC historically important is that it pioneered the 'precision first' approach to anaphora resolution that still underlies the best performing anaphora resolution systems, and, in particular, the Stanford Sieve approach discussed later in this Section [64, 48]. CogNIAC resolves pronouns by applying a series of rules ordered so that the most reliable (over a set of 200 'training' pronouns) apply first. Another sense in which CogNIAC is precision oriented is that its basic version does not attempt to resolve all pronouns, but only those to which rules of sufficient precision apply. The six rules, with their performance on the 'training' pronouns, are as follows:

1. **UNIQUE IN DISCOURSE**: if there is a single matching antecedent i in the read-in portion of the entire discourse, then pick i as the antecedent.
   Accuracy: 8 correct, 0 incorrect
2. **REFLEXIVE** Pick nearest possible antecedent in read-in portion of current sentence if the anaphor is a reflexive pronoun.
   Example of application: Mariana motioned for Sarah to seat herself on a two-seater lounge.
   Accuracy: 16 correct, 1 incorrect
3. **UNIQUE IN CURRENT + PRIOR** If there is a single possible antecedent i in the prior sentence and the read-in portion of the current sentence, then pick i as

the antecedent.
Example of application: Rupert Murdoch's News Corp. confirmed his interest in buying back the ailing New York Post. But analysts said that if he winds up bidding for the paper,....
Accuracy: 114 correct, and 2 incorrect

4. **POSSESSIVE PRO** If the anaphor is a possessive pronoun and there is a single exact string match i of the possessive in the prior sentence, then pick i as the antecedent.
Accuracy: 114 correct, and 2 incorrect
5. **UNIQUE CURRENT SENTENCE** If there is a single possible antecedent in the read-in portion of the current sentence, then pick i as the antecedent.
Accuracy: 21 correct, and 1 incorrect
6. **UNIQUE SUBJECT / SUBJECT PRONOUN** If the subject of the prior sentence contains a single possible antecedent i, and the anaphor is the subject of its sentence, then pick i as the antecedent.
Example of application: Besides, if he provoked Malek, uncertainties were introduced, of which there were already far too many. He noticed the supervisor enter the lounge ...
Accuracy: 11 correct, and 0 incorrect

In [5], CogNIAC was systematically evaluated on narrative texts (where its performance was compared with that of Hobbs' naive algorithm, finding similar performance), on WSJ texts (achieving a recall of 78% and precision of 89%), and over the 30 articles in the MUC-6 test data (CogNIAC was the pronoun resolution component of the University of Pennsylvania's MUC-6 submission) achieving a recall of 75% and a precision of 73%.

3.7.2 MARS

Mitkov's MARS, like CogNIAC, is based on the assumption that a great number of pronouns can be resolved using what Mitkov calls **knowledge-poor** methods ([54]; see also Chapter 7 of [56]). Specifically, MARS relies only on the output of a Part-of-Speech tagger and of a parser–Conexor's FDG dependency parser [36],also used by ROSANA.

And indeed, MARS can be viewed as a stripped-down version of ROSANA: choose as actual antecedent the one among the potential antecedents that matches the pronoun in gender and number and has the higher 'score'. More specifically, MARS consists of five steps:

1. Parse the text using the FDG parser, that extracts parts-of-speech, lemmas, syntactic function, number, and dependency relations between the NPs.
2. Identify the pronouns to be processed. MARS only attempts to resolve third person personal and possessive pronouns; non-anaphoric instances of *it* are identified using Evans' algorithm [13].

3. The **competing candidates** of every pronoun identified in phase 2 are extracted. These are the NPs in the current and preceding two sentences that match the pronoun in gender and number and pass three **syntax filters** derived from [44].
4. The **antecedent indicators** (14 in total) are applied to each potential candidate to compute its score.
5. The candidate with the highest score is chosen. If two candidates have an equal score, the most recent candidate is chosen.

The heart of MARS are the rules for calculating the antecedent indicators. These rules are heuristics expressing preferences deriving from syntax, lexical / commonsense knowledge, and salience, and can either increase ('boost') the score of a candidate antecedent or decrease ('impede') it. Examples of boosting indicators include:

- *First noun phrase*: this indicator increments the score of the first NP in a sentence by +1 - i.e., it aims to capture the first mention advantage (as discussed in Chapter 2).
- *Indicating verbs*: this indicator increases by +1 the score of NPs that immediately follow certain verbs–on the basis of evidence about the so-called **implicit causality** effect [17, 70].

Examples of impeding indicators are

- *Indefiniteness*: The score of indefinite NPs is decreased by 1 by this indicator, in keeping with evidence that definite NPs are more salient.

Referential distance is an example of an indicator that can either increase or decrease the score of a potential antecedent: antecedents preceding the pronoun but occurring in the same sentence have their score increased by +2, antecedents in the previous sentence by +1, antecedents in the sentence before that by 0, and all other antecedents have their score decreased by 1.

Different versions of MARS incorporating slightly different syntax filters and indicators were developed and evaluated for English, Arabic, Polish and Bulgarian. The English version was evaluated on a corpus of eight computer hardware / software technical manuals, containing a total of 247,401 tokens and 2263 anaphoric pronouns. The best success rate was 61.55%. The Bulgarian version was evaluated on texts from two different domains containing a total of 221 pronouns, achieving a success rate of 75.7%.

MARS has been very influential, and versions of the algorithm have also been incorporated in platforms such as GUITAR [57].

3.8 Definite descriptions: the Vieira and Poesio algorithm

Most of the approaches and algorithms described in the Chapter so far, whether theoretically inspired or heuristic-based, deal mainly or exclusively with pronominal anaphora; insofar as they cover noun phrases and proper nouns (e.g., in terms of Binding Theory: type C occurrences), the resolution heuristics employed (e. g., string matching) are quite simple and surface-oriented. (The two exceptions are Sidner's algorithm, which covers all definite noun phrases, and Hobbs' commonsense-knowledge based approach, which covers all noun phrases.) There are two reasons for this focus on pronouns: a theoretical one—pronominal anaphora is much more governed by grammatical competence that full nominal anaphora—and a practical one—interpreting pronouns depends less on lexical, commonsense and encyclopedic knowledge than other types of anaphoric interpretation; hence, shallow approaches are more likely to achieve good results for this type of anaphora. By contrast, Vieira and Poesio [78, 79, 82, 59, 58, 77, 80, 81] deliberately focused on definite descriptions in their research, as the type of nominal anaphora most likely to lead to interesting findings about the effect of lexical and commonsense knowledge on anaphoric interpretation. In contrast to pronouns, which only encode grammatical information and degree of salience, definite descriptions such as *the man*, or *the city* encode much more information. As such, they are often used to realize subsequent mentions for expressions that are less salient because they are farer away and/or because they are non-animate, and the choice of potential antecedents is far greater. In addition, definite noun phrases that are sufficiently informative (*the president of Peru*, *the man I met yesterday*) can be non-anaphoric/discourse-new (i.e., correspond to a newly introduced entity). In this Section, we discuss their system(s).

3.8.1 Corpus analysis

The system developed by Vieira and Poesio was the first anaphora resolution system based on a systematic corpus annotation (of around 1,400 definite descriptions in the WSJ portion of the Penn Treebank) [58, 77] employing a reliability analysis in the sense of [7]. The annotation was designed to identify the major classes of definite descriptions so as to plan the effort. Familiarity-based theories (e.g., [24]) would predict that the majority of definite descriptions would be anaphoric; this would suggest putting most of the time on improving the resolution of anaphoric definites, as indeed done by most systems previously. By contrast, uniqueness-based theories (e.g., [50, 62]) would view anaphoric definites as only one type of definite description, and not necessarily the main one.

Vieira and Poesio's corpus annotation provided support for a uniqueness-based analysis. Only between 30% and 40% of definite descriptions in the corpus could be considered discourse-old in the sense of Prince; between 60 and 70% were discourse-new, including larger-situation definites (*the Iran-Iraq war*), unfamiliar cases (*the result of the analysis is ...*) and associative descriptions (also known as bridging, see below). The analysis of bridging references carried out by Vieira, Teufel and Poesio [82, 59] also identified those cases of associative description that could be reliably identified and resolved using lexical resources such as WordNet [14].

As a result, the system(s) developed by Vieira and Poesio [79, 77, 80, 81] include three types of methods: a set of heuristics to determine whether a definite description is likely to be discourse-new; a second set of heuristics to determine if a noun phrase in the preceding text is likely to be the antecedent to the (suspected) anaphoric definite noun phrase; and finally, heuristics relying on WordNet to identify the possible **anchors** of bridging references. We discuss each type of method in turn.

3.8.2 Heuristics for recognizing discourse-new definite descriptions

Discourse-new descriptions include, first of all, those definites that [23] called **larger situation uses**–terms whose uniqueness can be established on the basis of encyclopedic knowledge, such as *the pope*, *the moon*, *the sky*, or terms that have only one, or only one salient, referent for their class, such as time references (*hour*, *time*, *month*). In general, recognizing such cases requires encyclopedic knowledge; Vieira and Poesio's system used instead a series of heuristics. First of all, the system used a small list of such terms. Second, the system included a heuristic to classify as larger situation definites whose modifiers included a named entity (*the Iran-Iraq war*) or a numerical modifier (*the 1987 stock-market crash*).

Both [50] and [23] also discussed however a number of additional categories of discourse-new definites that could be recognized without recourse to encyclopedic knowledge. A first example are superlatives such as *the richest* and other definites modified by ordinals such as *first* or by modifiers such as *only* or *best*. The use of such definites presupposes a unique first (only, or best) element of a given universe, as in

(3.10) Mr. Ramirez just got *the first raise he can remember in eight years*, to $8.50 an hour from $8.

[23] grouped such definites into a class of definites relying on **special predicate** premodifiers, and Vieira and Poesio's system included a heuristic that suggested that definites with one of those modifiers are discourse-new.

Another category of non-large situation discourse-new definites are those whose head is a **functional predicate with complement**. This includes predicates such as *result*, *fact*, or *idea*, which are interpreted as nonanaphoric when they have a complement (typically a clause):

(3.11) Mr. Dinkins also has failed to allay Jewish voters' fears about his association with the Rev. Jesse Jackson, despite *the fact that few local non-Jewish politicians have been as vocal for Jewish causes in the past 20 years as Mr. Dinkins has.*

Vieira and Poesio's system included a list of such functional predicates.

Finally, Vieira and Poesio's system include syntax-based methods to recognize postmodifiers typical of discourse-new noun phrases, particularly **restrictive postmodifiers** and **appositions**. As pointed out by Hawkins, definites modified by relative clauses which are introduced by relative pronouns (such as *who, whom, which, where, when, why*, or *that*) are typically discourse new. Other indicators of discourse novelty are non-finite clauses or prepositional phrases when they occur as postmodifiers.

Much subsequent research has been carried out exploring discourse new detection; this research is discussed in a separate Chapter, Chapter 13.

3.8.3 Anaphoric definites, same-head antecedent

The simplest heuristic in the resolution of definite description is to look for a potential antecedent which has the same head as the head of the definite description (e.g., *a mushroom–the mushroom*). Vieira and Poesio's system includes additional heuristics to improve the precision of the resolution.

Firstly, **segmentation heuristics** filter out potential antecedents that are not salient enough or too far away – corresponding to the intuition that a referent introduced in the second sentence of an article may have been forgotten when the reader is at the 50th sentence of the article. In its simplest form, such a heuristic may filter out any potential antecedents that are more than *n* sentences away. However, such a hard constraint filters out a considerable number of correct antecedents. Vieira and Poesio therefore use additional criteria in their **loose segmentation** heuristic. This heuristic admits potential antecedents from outside the given window (typically 4 sentences) if the antecedent is either (i) a subsequent mention (corresponding to the intuition that subsequent mentions are somehow entrenched and do not fade out of memory as quickly) or (ii) string-identical (including the article) to the previous mention.

Secondly, **compatibility heuristics** attempt to deal with the fact that noun phrases with the same noun head are sometimes not valid antecedents when they have incompatible modifiers: For example, *a blue car* is not a valid antecedent for *the red car*, or that *software from the US* and *the software from India* cannot co-refer.

Vieira and Poesio's heuristics consider both premodifiers and postmodifiers, and generally treat modifiers as incompatible when they have different surface strings. Their heuristic admits certain kinds of subset/superset relations for the modifiers of an antecedent:

- When the premodifiers of the antecedent are a superset of the premodifiers of the definite description (e.g. *the colored car* to *a* <u>*blue*</u> *car*)
- When a possible antecedent has no premodifiers at all (in which case the additional premodifiers of the definite description are assumed to include new information about an old referent, as in resolving *the* <u>*lost*</u> *check* to *a check*).

If there are multiple potential same-head antecedents, the closest one is chosen (**recency heuristic**).

3.8.4 Bridging descriptions

The antecedent of many anaphoric definite descriptions has a nominal head that differs from that of the definite description. This is one example of bridging reference [11]. Vieira and Poesio's system includes methods for dealing with this class of bridging references (for which Vieira and Poesio use the term **coreferent bridging**) as well as for some types of **associative bridging**. In their corpus, about 15% of definite descriptions belong to their "bridging" category, against about 30% of all definite descriptions that have a same-head antecedent.

Some cases of coreferent bridging depend on lexical knowledge about lexical or conceptual relations that is available from WordNet. This includes the case in which the two head nouns are synonyms of each other (*suit* vs. *lawsuit*), or when the head noun of the antecedent is a hyponym of the head noun of the anaphor (as in *dollar* vs. *currency*). In the case of near-synonymy, the synsets may also be coordinate sisters, i.e., direct hyponyms of a common hyperonym (e.g., *home* and *house*).

Another heuristic for coreferent bridging proposed by Vieira and Poesio concerns references to **named entities**, as in *Pinkerton Inc.* subsequently mentioned as *the company*. To recognize such links, Vieira and Poesio use a combination of a named entity recognizer (which detects named entities and categorizes them as either person, location, or organization) and the conceptual knowledge in WordNet. Coreference is assumed when the definite description is a hyponym of a synset that is indicative for one of the named entity categories (in particular, *country*, *city*, *state*, *continent*, *language* or *person*), and the postulated antecedent has this particular named entity type. To aid this process, the output of the named entity recognizer is refined by using full name mentions such as *Mr. Morishita* to assign a named entity type to ambiguous shortened mentions such as *Morishita*.

In cases of associative bridging, one heuristic relies on **meronymy** relations in WordNet, such as *the living room* being a part of *the flat*.

In several cases, a bridging description that is not introduced by a noun phrase can be recovered from syntactic material, such as cases in which a bridging antecedent is a **prenominal modifier** of the antecedent (*the discount packages* as antecedent to *the discounts*), but also cases in which the antecedent to a definite description is a **verb phrase** such as the clause *Kadane oil is currently drilling two oil wells* licensing a subsequent definite noun phrase *the activity*.

Yet other cases, such as definite descriptions that are licensed through discourse topic or other means (as in *the industry* in a text referring to oil companies) or more general world knowledge, are completely out of the reach of heuristics such as those proposed by Vieira and Poesio both because they present challenges for the annotation [58] and because the necessary lexical / commonsense knowledge was not available.

3.8.5 Putting it all together

The overall architecture of the Poesio / Vieira system is very simple. The system goes through the text sentence by sentence. Whenever a new sentence is encountered, the segmentation window is updated, and all mentions extracted.[10] The system then heuristically identifies all definite descriptions; all the other NPs, except for pronouns, are taken to be discourse-new, and to introduce a new **file card** (the internal representation of discourse entities). The system then uses a decision-tree (hand-coded or learned, see below) to classify each definite description as discourse-new or discourse-old using the heuristics; a new file card is created for every definite description classified as discourse-new, whereas the information for discourse-old entities is added to the file card of their antecedent.

Both the selection of the specific heuristics to be included in the decision tree, and the overall order of the heuristics, were carried out empirically, using a corpus of about 1,000 annotated definite descriptions for development, and a test set consisting of about 400 definite descriptions for testing. We summarize some of results here; see [77] for details.

Choosing among different variants of a heuristics

Vieira and Poesio found that their loose segmentation with recency heuristic (81.44% F for a loose 4-sentence window) works considerably better than only using recency (79.62% F), which in turn works better than just using a hard distance cutoff (69.76% F for a strict window).

In the realm of compatibility heuristics, they show that the system using the two heuristics (requiring that the antecedent either contains a superset of the definite description's modifiers, or has none at all) works better (with, again, 81.44% F) than a version that does not check modifier compatibility (80.19% F) or that only allows antecedents with a superset of the definite description's modifiers (79.12% F).

The heuristics for first-mention uses of definite descriptions show varying precision ranging from 75-93% for most heuristics (postmodification, apposition, names, time references, unexplanatory modifiers) on the training data. Vieira and Poesio identify problem sources in copula construction (where the distinction between subject and predicate in sentences such as 3.12 is not always clear), and in restrictive premodification, which can also carry new information about an old antecedent:

(3.12)a. *The key man* seems to be the campaign manager, Mr. Lynch.
b. in the fear that an aftershock will jolt *the house* again.
...
As Ms. Johnson stands outside *the <u>Hammock</u> house* after winding up her chorse there, the house begins to creak and sway.

[10] Sentences and mentions are gold, extracted from the Penn Treebank annotation. The mentions and heuristically aligned with the output of a NE recognizer.

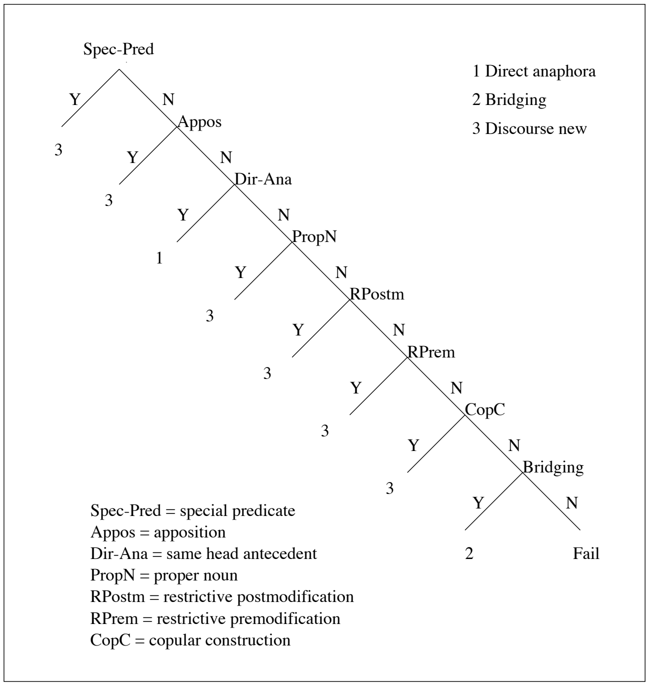

Spec-Pred
Y
N
3
Appos
Y
N
3
Dir-Ana
Y
N
1
PropN
Y
N
3
RPostm
Y
N
3
RPrem
Y
N
3
CopC
Y
N
3
Bridging
Y
N
2
Fail
1 Direct anaphora
2 Bridging
3 Discourse new
Spec-Pred = special predicate
Appos = apposition
Dir-Ana = same head antecedent
PropN = proper noun
RPostm = restrictive postmodification
RPrem = restrictive premodification
CopC = copular construction

Fig. 3.4 Hand-coded decision tree of Version 1

In the realm of bridging descriptions, WordNet relations generally have a precision ranging from 36% (synonymy) to 20% (coordinate sisters). Vieira and Poesio conclude that the knowledge encoded in WordNet is not sufficient to interpret all semantic relations that are involved in bridging resolution.

Optimal configuration of the heuristics

Vieira and Poesio ended up with seven heuristics for identifying discourse-new definites and resolving same-head anaphoric definites, as well as additional heuristics for resolving bridging descriptions. These heuristics were employed to develop three variants of their system, all of which based on decision trees:

1. A version ignoring bridging references and only attempting to identify discourse-new and discourse-old descriptions and to resolve the latter, in which the decision tree specifying the order of the heuristics was determined by their precision, as in CoGNIAC (see Figure 3.4). (This version is called ’Version 1’ in [81].)
2. A second version including the methods for resolving bridging references, called Version 2 in the paper;

3. A third version also ignoring bridging references, but in which the decision tree specifying the order of the heuristics was determined using the ID3 decision tree learning algorithm [63], using the development corpus as training data.

The performance of the three versions is compared in Table 3.1. As shown in the Table, Version 1 of the algorithm with the hand-coded decision tree achieves an overall F of .62. Version 2 achieves the same overall F, but with a higher precision and a lower recall. Version 1 with an automatically learned decision tree achieves a much higher F (.75) as it assigns a DN classification to all NPs for which no other rule applies. The best overall results (F = .77(are obtained by a variant of the hand-coded form of Version 1 that also automatically classifies as DN all definites that haven't been given any other classification.

Version	**P**	**R**	**F**
Version 1 (hand-coded)	73	56	62
] Version 2	70	57	62
Version 1 (ID3)	75	75	75
Version 1 (hand-coded+DN default)	**77**	**77**	**77**

Table 3.1 Comparison between the three versions of the Vieira and Poesio system.

Examining the differences between the hand-coded and automatically learned decision tree, Vieira and Poesio found that the only difference was in the very first test: whereas the hand-coded version starts by checking whether the NP contains special premodifiers, the automatically learned version starts by checking if a same-head antecedent exists. This is especialy interesting at the light of the subsequent work on non-anaphoricity detection discussed in Chapters 9, 9.5 and 13.

3.9 Rule-based and Heuristic Systems in the MUC 6 and MUC 7 Coreference Task

The approaches and systems presented in the earlier sections of this Chapter are for the most part focused on a specific type of anaphoric reference– pronouns for the historical proposals by Hobbs, Charniak, the algorithms based on centering, and more recent algorithms such as Lappin and Leass' RAP, MARS, and CoGNIAC; and definite descriptions in the case of Vieira and Poesio's resolution approach. This narrow focus was motivated by evidence highlighting how differently each type of anaphoric expression behaves both from a linguistic and from a processing perspective; but the result were systems unable to handle the complexity of full anaphoric reference. This all changed with the 6th and 7th edition of the Message Understanding Conferences, where a **Coreference (CO) Task** was introduced [20, 26]. In order to achieve a high performance in those evaluation campaigns, an *integrated account* of pronominal and definite description resolution, discourse-new detection, proper name recognition / classification / matching, and a differential treatment of text regions turns out to be required, thus making this problem considerably harder than the mere implementation of a high-accuracy pronoun resolver.

In this section, we will discuss first the coreference component of the FASTUS system [3] , developed by Kameyama [38]. Kameyama's system was the best performing system in the MUC-6 evaluation, with a MUC F-measure of 0.65.[11] Next, we will discuss the coreference component of the LaSIE II system. The LaSIE systems [15, 35] participated at both MUC-6 and MUC-7. LaSIE-II was the best-performing system in the MUC-7 coreference task, with an F-measure of 0.618.[12]

In having triggered the development of these important full-fledged coreference resolvers, the MUC CO Task Evaluations can be regarded as *the* decisive momentum towards resources-driven, robust anaphora processing.

3.9.1 The FASTUS system by Kameyama et al.

The key characteristic of the rule-based system developed by [38] for MUC-6 is that it builds on the finite-state grammar developed for the FASTUS system, versions of which participated in several editions of MUC [3]. FASTUS yields a chunking analysis of the text which is highly accurate in the presence of complex noun chunks, but does not produce the type of hierarchical structure that, e.g. Hobbs' or Lappin and Leass' approaches for pronoun resolution or the approach of Vieira and Poesio for definite descriptions, presuppose. The anaphora resolution system described in [38], therefore, approximates appositional/copular constructions and (originally syntactic) salience within the pattern-based approach in FASTUS. Kameyama points out that these approximations lead to a loss of precision with respect to perfect or good parses used in other systems; however, the loss due to this approximation approach is not as large as one could imagine, and the most obvious cases where a syntactic analysis would help (reflexives and disjoint reference filtering) are relatively infrequent.

3.9.1.1 Mention detection

Kameyama's system takes mentions (**template entities**) as input which, besides their span, already have some linguistic features that are useful for subsequent processing:

- the determiner or pronoun type (definite, indefinite, or pronominal);

[11] Soon et al.'s system [69], the first successful machine learning approach, discussed in Chapter 9, obtained an F score of 0.63 for this dataset. As we will see in the rest of this Chapter and in the following Chapters of the book, it is still the case for coreference that a rule-based system can achieve state-of-the-art performance.

[12] Soon et al.'s system obtained an F of 0.605.

- grammatical number (singular or plural, or a modifying cardinal expression);
- head string and modifiers of the mention's noun chunk;
- a semantic class that is assigned based on the head, and comes from a shallow hierarchy;
- sentence and paragraph positions;
- information about the enclosing text region (headline or main text);

Information about the enclosing **text region** is used to model the assumptions for text-region accessibility that were employed in the MUC-6 annotation, namely that a mention in the *headline* region can be coreferring with a mention from the text, whereas mentions in the *text* region can be resolved to any preceding mention within the *text* region.

3.9.1.2 Resolution strategy

The system has different resolution strategies based on the type of the mention, using different heuristic constraints for pronouns, definite descriptions, and names.

For **pronouns**, a narrow three-sentence window is enforced (respectively, only the current sentence is used for reflexives), together with consistency constraints for *number* and *semantic sort*, and subsequently ranked based on a left-right ranking order (see below).

Plural pronouns (*they*, *we*) are considered consistent with singular organization antecedents. First person pronouns (*I*, *we*) are allowed to be resolved as intrasentential cataphora (i.e., to a later mention in the same sentence), unlike non-pronouns or other types of pronouns.

The resolution of **definite noun phrases** relies on a window size of ten sentences, together with a *sort consistency* constraint that requires the sort of the anaporic definite noun phrase to be equal or more general than the sort of the antecedent candidate. This would allow *the company* as a subsequent mention of *the automaker* but not vice versa. Similar in spirit to Vieira and Poesio's approach, anaphor-antecedent pairs with known-inconsistent modifiers (for example *French* and *British*) are filtered out.

For **proper names**, the entire previous text is considered. The semantic class of names is provided by FASTUS' heuristics for recognizing specific-type names (persons, locations or organizations), as well as unknown names.

The system considers both *shortened names* (alias) which have a selective substring of the full name (e.g. *Colonial* for *Colonial Beef*), but also *acronyms* which have a subsequence of the initial characters of a name (e.g., *GM* for *General Motors*). For names with an unknown semantic class, the *merging of entities* in the resolution process makes sure that previously unknown aliases of entities for which the semantic class is known also get the semantic class information from their previous mention.

The **salience ordering** used by Kameyama's system is similar to the Left-Right Centering approach for pronouns as proposed by [76] in that it uses sentence information and surface order, but not the kind of hierarchical syntactic structure that earlier approaches would require.

Here, the *preceding part of the same sentence* is ordered left to right (i.e., topicalized phrases and subjects first), followed by the *immediately preceding* sentence, also in left-to-right order. Other preceding sentences (up to the resolution window, which depends on the expression type of the mention).

3.9.1.3 Evaluation

Overall, Kameyama's system scored 59% recall and 72% precision (F=.65) in the official MUC-6 evaluation, which was the best overall performance [74]. In the 1997 article Kameyama also includes a more detailed breakdown of the performance by the type of mention. Intra-sentential 3rd person pronouns (27 mentions) are resolved with 78% precision, whereas inter-sentential 3rd person pronouns (33% precision, 6 mentions) and 1st/2nd person pronouns (20% precision, 5 mentions) are more difficult. The system achieves a precision of 69% in the resolution of proper names (32 resolved mentions), and a considerably lower figure of 46% for definite descriptions (61 occurrences).[13]

3.9.2 Coreference in the LaSIE systems

The LaSIE [15] and LaSIE-II [35] systems, developed at the University of Sheffield, participated in MUC-6 and MUC-7, respectively.

The systems share the same broad approach. They are both implemented using a very modular and very general pipeline architecture not focused solely on the MUC tasks but including all the traditional components of an NLP system, from tokenizer to POS tagger to (statistical) parser to NE tagger to semantic interpreter including wordsense disambiguation to a coreference component. One of the main differences between LaSIE and LaSIE-II is that the pipeline used for LaSIE became the basis for the GATE NLP platform, which in turn became the basis for LaSIE-II. The systems also share the basis philosophy, aptly described by Humphreys et al [35] as threading "a pragmatic middle way between shallow vs deep analysis" resulting in the employment of "an eclectic mixture of techniques".

From an anaphora resolution perspective, the most significant aspect of the LaSIE systems is that they attempt to build a full **discourse model** in the sense advocated by linguists and psychologists [39, 16]: i.e., these systems not only (i) attempt to link every mention to an existing discourse entity, or to create a new one otherwise, as done as well by the Vieira / Poesio system; but they also (ii) attempt to expand the

[13] This figure cannot be compared to the figures obtained by Vieira and Poesio, because the latter evaluate the *resolution accuracy* for definite descriptions, whereas Kameyama's evaluation requires both correct identification of a discourse-old noun phrase and the identification of the correct antecedent to be counted.

bare-bones model consisting of these discourse entities into a proper domain model using an ontology.

The coreference component of LaSIE II is for the most part is a incremental development of the system incorporated in LaSIE-I, but it extends the earlier system to include

- look for antecedents not just in the current and previous paragraph but also further away–according to Humphreys et al this resulted in a 2% increase in recall with no significant effect on precision;
- methods for 'resolving' mentions occurring in copula constructions such as *the Navy's first-line fighter* in *The F-14 "Tomcat" is the Navy's first-line fighter* which, as discussed in Chapters 2 and 4, are treated as cases of 'coreference' in MUC;
- methods for resolving some cases of cataphora, and 'bare noun' coreference;
- methods for resolving som cases of NPs occurring in coordination (e.g., *John* and *his boys* in *John and his boys* introduce new discourse entities.

Both systems performed very well at the coreference task. LaSIE-I achieved a R = 0.51, P = 0.71, F = .59 at MUC-6 (third highest) whereas LaSIE-II was the best system at the coreference task in MUC-7, obtaining R=.56, P=.69, F=.62.

3.10 Modern Heuristic-Based Approaches: The Stanford Deterministic Coreference Resolution System

The heuristic approach to the development of coreference systems is thriving. Many such systems are still being developed; indeed, the Stanford Deterministic Coreference Resolution System,[14] based on the so-called 'Stanford Sieve' approach [64, 47, 48]– a version of the 'precision-first' approach pioneered by CogNIAC and also adopted by MARS and the Vieira-Poesio system, was the best performing system at the CoNLL 2011 coreference shared task [47]; at CoNLL 2012, two of the three best-performing systems (namely, Fernandes et al. **[REF FIXME]** as well as Chen and Ng **[REF FIXME]**) were hybrid models that used machine learning on top of the resolutions of the Sieve model.[15]

In fact, these modern, hybrid approaches perfectly illustrate the eclecticism that is prevailing in current research, thus evidencing that it is justified to speak of the current research period, as suggested in Chapter 1, as the post-modern phase.

The architecture of the Stanford DCR is articulated around two main stages: a high recall (and highly precise) mention detection component based on Stanford CoreNLP, a high quality NLP pipeline[16]; and a coreference resolution stage consisting of 10 components called (**sieves**) analogous to CoGNIAC's rules and also

[14] `http://nlp.stanford.edu/software/dcoref.shtml`

[15] The CoNLL coreference shared tasks are discussed in detail in Chapter 6.

[16] `http://nlp.stanford.edu/software/corenlp.shtml`

ordered from the highest precision to lowest precision. The operation of the coreference resolution stage is based on the following principles:

- The system keeps track of **entities** (i.e., the discourse entities of systems such LaSIE: sets of mentions that have already been determined to belong together), while keeping track of properties such as number, gender, animacy, and named entity type.
- Each sieve operates on entities rather than mentions, and on the whole discourse, rather than on a sentence or a paragraph at a time.
- The system also keeps track of **cannot-link** constraints that have been added at various steps – i.e., for two entities, components can both add **must-link** constraints (merge the entities) or **cannot-link** constraints (such that the entities cannot be merged by any downstream component).
- For sieves that compare two mentions, the system keeps track of a "representative" mention in each cluster (typically the first one, as it is usually the longest, whereas subsequent mentions are shortened or only expressed as pronouns).

The ten sieves are:

1. *Speaker Identification*: This sieve first identifies **speakers**, then matches first and second pronouns to these speakers.
2. *Exact Match*: This sieve links together two mentions only if they contain exactly the same text, including both determiners and modifiers.
3. *Relaxed String Match*: This sieve links together two mentions only if they contain exactly the same text after dropping the postmodifiers.
4. *Precise Constructs*: This sieve links together two mentions if they occur in one of a series of high precision constructs: e.g., if they are in an appositive construction (*[the speaker of the House], [Mr. Smith]* ...), or if both mentions are tagged as NNP and one of them is an acronym of the other.
5. *Strict Head Match*: This sieve links together a mention with a candidate antecedent entity if *all* of a number of constraints are satisfied: (a) the head of the mention matches any of the heads of the candidate antecedent; (b) all non-stop words of the mention are included in the non-stop words of the candidate antecedent; (c) all mention modifiers are included among the modifiers of the candidate antecedent; and (d) the two mentions are not in an i-within-i situation, i.e., one is not a child in the other.
6. *Variants of Strict Head Match*: Sieve 6 relaxes the 'compatible modifiers only' constraint in the previous sieve, whereas Sieve 7 relaxes the 'word inclusion' constraint.
7. *Proper Head Match*: This sieve links two proper noun mentions if their head words match and a few other constraints apply.
8. *Relaxed Head Match*: This sieve relaxes the requirement that the head word of the mention must match a head word of the candidate antecedent entity.
9. *Pronoun resolution*: Finally, pronouns are resolved, by finding candidates matching the pronoun in number, gender, person, animacy, and NER label, and at most 3 sentences distant.

The Stanford Deterministic Coreference System achieved the highest MELA score (59.5) at the CONLL 2011 coreference shared task; and has been extensively evaluated on a variety of other datasets, always achieving state-of-the-art results.

3.11 Conclusions

In this Chapter we have covered in some detail most of the best-known non-statistical approaches to anaphora resolution. As seen discussing the Stanford Deterministic Coreference System, such approaches still achieve state-of-the-art performance, and very few new ideas about the linguistic features playing a role in anaphora resolution have been introduced in more recent systems; but the thrust of the research in the field has moved towards statistical methods. The remaining chapters will focus on these approaches.

References

[1] Alshawi H (1987) Memory and Context for Language Interpretation. Cambridge University Press, Cambridge

[2] Alshawi H (ed) (1992) The Core Language Engine. The MIT Press

[3] Appelt DE, Hobbs JR, Bear J, Israel D, Kameyama M, Tyson M (1993) Fastus: A finite-state processor for information extraction from real-world text. In: Proc. IJCAI, Chambery

[4] Asher N, Lascarides A (2003) The Logic of Conversation. Cambridge University Press

[5] Baldwin B (1997) Cogniac: A high precision pronoun resolution engine. In: Proceedings of the ACL97/EACL97 workshop on Operational Factors in Practical, Robust Anaphora Resolution, Madrid, pp 38–45

[6] Brennan S, Friedman M, Pollard C (1987) A centering approach to pronouns. In: Proc. of the 25th ACL, Stanford, CA, pp 155–162

[7] Carletta J (1996) Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics 22(2):249–254

[8] Carter DM (1987) Interpreting Anaphors in Natural Language Texts. Ellis Horwood, Chichester, UK

[9] Charniak E (1972) Towards a model of children's story comprehension. PhD thesis, MIT, available as MIT AI Lab TR-266

[10] Charniak E (1975) Organization and inference in a frame-like system of commonsense knowledge. In: Proc. of TINLAP, pp 42–51

[11] Clark HH (1975) Bridging. In: Schank RC, Nash-Webber BL (eds) Proceedings of the 1975 workshop on Theoretical issues in natural language processing, Association for Computing Machinery, Cambridge, MA, pp 169–174

[12] Dale R (1992) Generating Referring Expressions. The MIT Press, Cambridge, MA

[13] Evans R (2001) Applying machine learning toward an automatic classification of it. Literary and Linguistic Computing 16(1):45–57

[14] Fellbaum C (ed) (1998) WordNet: An electronic lexical database. The MIT Press

[15] Gaizauskas R, Wakao T, Humphreys K, Cunningham H, Wilks Y (1995) University of Sheffield: description of the LaSIE System as used for MUC-6. In: Proceedings of the Sixth Message Understanding Conference (MUC-6), Morgan Kauffmann, pp 207–220

[16] Garnham A (2001) Mental models and the interpretation of anaphora. Psychology Press

[17] Garvey C, Caramazza A (1974) Implicit causality in verbs. Linguistic Inquiry 5:459–464

[18] Ge, N., Hale, J., Charniak, E.: A statistical approach to anaphora resolution. In: Proc. WVLC/EMNLP 1998 (1998)

[19] Gordon PC, Grosz BJ, Gillion LA (1993) Pronouns, names, and the centering of attention in discourse. Cognitive Science 17:311–348

[20] Grishman R, Sundheim B (1995) Design of the MUC-6 evaluation. In: Proceedings of the Sixth Message Understanding Conference (MUC-6)

[21] Grosz BJ (1977) The representation and use of focus in dialogue understanding. PhD thesis, Stanford University

[22] Grosz BJ, Joshi AK, Weinstein S (1995) Centering: A framework for modeling the local coherence of discourse. Computational Linguistics 21(2):202–225, (The paper originally appeared as an unpublished manuscript in 1986.)

[23] Hawkins J (1978) Definiteness and Indefiniteness. Croom Helm, London

[24] Heim I (1982) The semantics of definite and infedinite noun phrases. PhD thesis, University of Massachusetts at Amherst

[25] Hewitt C (1969) Planner: a language for proving theorems in robots. In: Proc. of IJCAI, pp 295–302

[26] Hirschman L, Chinchor N (1997) MUC-7 coreference task definition (version 3.0). In: Proceedings of the 7th Message Understanding Conference, URL `http://www-nlpir.nist.gov/related\_projects/muc/proceedings/co\_task.html`

[27] Hirst G (1981) Discourse-oriented anaphora resolution: A review. Computational Linguistics 7:85–98

[28] Hobbs J (1978) Resolving pronoun references. Lingua 44:311–338

[29] Hobbs JR (1976) Pronoun resolution. Research Note 76-1, City College, City University of New York

[30] Hobbs JR (1979) Coherence and coreference. Cognitive Science 3:67–90

[31] Hobbs JR (1986) Discourse and inference, unpublished draft

[32] Hobbs JR, Martin P (1987) Local pragmatics. In: Proc. IJCAI-87, Milano, Italy, pp 520–523

[33] Hobbs JR, Appelt DE, Bear J, Tyson M, Magerman D (1991) The tacitus system: The muc-3 experience. SRI Technical Note 511, SRI International, Menlo Park, CA

[34] Hobbs JR, Stickel M, Appelt D, Martin P (1993) Interpretation as abduction. Artificial Intelligence 63:69–142

[35] Humphreys K, Gaizauskas R, Azzam S, Huyck C, Mitchell B, Cunningham H, Wilks Y (1998) University of sheffield: Description of the lasie-ii system as used for muc-7. In: Proceedings of MUC-7

[36] Järvinen T, Tapanainen P (1997) A dependency parser for english. Technical Report TR-1, Department of General Linguistics, University of Helsinki

[37] Kameyama M (1985) Zero anaphora: The case of japanese. PhD thesis, Stanford University, Stanford, CA

[38] Kameyama M (1997) Recognizing referential links: an information extraction prespective. In: ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts

[39] Kamp H, Reyle U (1993) From Discourse to Logic. Kluwer, Dordrecht

[40] Kantor RN (1977) The management and comprehension of discourse connection by pronouns in english. PhD thesis, Ohio State University, Department of Linguistics

[41] Karamanis N, Poesio M, Oberlander J, Mellish C (2009) Evaluating centering for information ordering using corpora. Computational Linguistics 35(1):29–46

[42] Kehler A, Kertz L, Rohde H, Elman J (2008) Coherence and coreference revisited. Journal of Semantics 25(1):1–44

[43] Kennedy C, Boguraev B (1996) Anaphora for everyone: Pronominal anaphora resolution without a parser. In: COLING 1996

[44] Kennedy C, Boguraev B (1996) Anaphora for everyone: Pronominal anaphora resolution without a parser. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING), pp 113–118

[45] Kibble R, Power R (2000) An integrated framework for text planning and pronominalization. In: Proc. of the International Conference on Natural Language Generation (INLG), Mitzpe Ramon, Israel

[46] Lappin S, Leass H (1994) An algorithm for pronominal anaphora resolution. Computational Linguistics 20(4):535–561

[47] Lee H, Peirsman Y, Chang A, Chambers N, Surdeanu M, Jurafsky D (2011) Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In: Proc. of the CoNLL 2011 Shared Task

[48] Lee H, Chang A, Peirsman Y, Chambers N, Surdeanu M, Jurafsky D (2013) Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics 39(4):885–916

[49] Linde C (1979) Focus of attention and the choice of pronouns in discourse. In: Givon T (ed) Syntax and Semantics 12, Academic Press

[50] Löbner S (1985) Definites. Journal of Semantics 4:279–326

[51] Matthews A, Chodorow MS (1988) Pronoun resolution in two-clause sentences: effects of ambiguity, antecedent location, and depth of embedding. Journal of Memory and Language 27:245–260

[52] Miltsakaki E (2002) Towards an aposynthesis of topic continuity and intrasentential anaphora. Computational Linguistics 28(3):319–355

[53] Minsky M (1975) A framework for representing knowledge. In: Winston PH (ed) The Psychology of Computer Vision, McGraw-Hill, New York, pp 211–277

[54] Mitkov R (1998) Robust pronoun resolution with limited knowledge. In: Proc. of the 18th COLING, Montreal, pp 869–875

[55] Mitkov R (2002) Anaphora Resolution. Longman

[56] Mitkov R (2005) Anaphora resolution. In: Oxford Handbook of Computational Linguistics, Oxford University Press

[57] Poesio M, Kabadjov MA (2004) A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In: LREC 2004, URL `http://privatewww.essex.ac.uk/~malexa/html\_files/files/LREC2004.pdf`

[58] Poesio M, Vieira R (1998) A corpus-based investigation of definite description use. Computational Linguistics 24(2):183–216

[59] Poesio M, Vieira R, Teufel S (1997) Resolving bridging descriptions in unrestricted text. In: ACL-97 Workshop on Operational Factors in Practical, Robust, Anaphora Resolution For Unrestricted Texts

[60] Poesio M, Stevenson R, Di Eugenio B, Hitzeman JM (2004) Centering: A parametric theory and its instantiations. Computational Linguistics 30(3):309–363

[61] Prince EF (1981) Toward a taxonomy of given-new information. In: Cole P (ed) Radical Pragmatics, Academic Press, New York, pp 223–256

[62] Prince EF (1992) The ZPG letter: subjects, definiteness and information-status. In: Thompson S, Mann W (eds) Discourse description: diverse analyses of a fund raising text, John Benjamins B.V., Amsterdam

[63] Quinlan JR (1986) Induction of decision trees. Machine Learning 1(1):81–106

[64] Raghunathan K, Lee H, Rangarajan S, Chambers N, Surdeanu M, Jurafsky D, Manning C (2010) A multi-pass sieve for coreference resolution. In: Proc. EMNLP, MIT, Boston, pp 492–501

[65] Reichman R (1985) Getting Computers to Talk Like You and Me. The MIT Press, Cambridge, MA

[66] Sanford AJ, Garrod SC (1981) Understanding Written Language. Wiley, Chichester

[67] Sidner CL (1979) Towards a computational theory of definite anaphora comprehension in english discourse. PhD thesis, MIT

[68] Sidner CL (1983) Focusing in the comprehension of definite anaphora. In: Brady M, Berwick R (eds) Computational Models of Discourse, MIT Press, Cambridge, MA

[69] Soon WM, Ng HT, Lim DCY (2001) A machine learning approach to coreference resolution of noun phrases. Computational Linguistics 27(4):521–544, URL `http://acl.eldoc.ub.rug.nl/mirror/J/J01/J01-4004.pdf`

[70] Stevenson RJ, Crawley RA, Kleinman D (1994) Thematic roles, focus, and the representation of events. Language and Cognitive Processes 9:519–548

[71] Strube M (1998) Never look back: An alternative to centering. In: Proc. of COLING-ACL, Montreal, pp 1251–1257

[72] Strube M, Hahn U (1999) Functional centering–grounding referential coherence in information structure. Computational Linguistics 25(3):309–344

[73] Stuckardt R (2001) Design and enhanced evaluation of a robust anaphor resolution algorithm. Computational Linguistics 27(4):479–506

[74] Sundheim BM (1995) Overview of the results of the MUC-6 evaluation. In: Proc. of the Sixth Message Understanding Conference (MUC-6), Columbia, Maryland, pp 13–31

[75] Suri LZ, McCoy KF (1994) RAFT/RAPR and centering: A comparison and discussion of problems related to processing complex sentences. Computational Linguistics 20(2):301–317

[76] Tetrault J (2001) A corpus-based evaluation of centering and pronoun resolution. Computational Linguistics 27(4):507–520

[77] Vieira R (1998) Definite description resolution in unrestricted texts. PhD thesis, University of Edinburgh, Centre for Cognitive Science

[78] Vieira R, Poesio M (1996) Corpus-based approaches to NLP: a practical prototype. In: Anais do XVI Congresso da Sociedade Brasileira de Computa cão

[79] Vieira R, Poesio M (1997) Processing definite descriptions in corpora. In: Botley S, McEnery M (eds) Corpus-based and Computational Approaches to Discourse Anaphora, UCL Press

[80] Vieira R, Poesio M (2000) Corpus-based development and evaluation of a system for processing definite descriptions. In: Proc. of 18th COLING, Saarbruecken

[81] Vieira R, Poesio M (2000) An empirically based system for processing definite descriptions. Computational Linguistics 26(4):539–593

[82] Vieira R, Teufel S (1997) Towards resolution of bridging descriptions. In: ACL-EACL 1997

[83] Vilain M, Burger J, Aberdeen J, Connolly D, Hirschman L (1996) A model-theoretic coreference scoring scheme. In: Proceedings of the 6th Message Understanding Conference (MUC-6), Morgan Kaufmann, San Francisco, pp 45–52, DOI http://dx.doi.org/10.3115/1072399.1072405

[84] Walker MA (1989) Evaluating discourse processing algorithms. In: Proc. of ACL, pp 251–261

[85] Walker MA, Iida M, Cote S (1994) Japanese discourse and the process of centering. Computational Linguistics 20(2):193–232

[86] Walker MA, Joshi AK, Prince EF (eds) (1998) Centering Theory in Discourse. Clarendon Press / Oxford

[87] Wilks YA (1975) An intelligent analyzer and understander of english. Commun ACM 18(5):264–274, reprinted in *Readings in Natural Language Processing*, Morgan Kaufmann

[88] Wilks YA (1975) A preferential pattern-matching semantics for natural language. Artificial Intelligence Journal 6:53–74

[89] Winograd T (1972) Understanding Natural Language. Academic Press