In: Proc. 5th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC04), São Miguel/Azores, Sept. 2004, 157-163.

Three Algorithms for Competence-Oriented Anaphor Resolution

Roland Stuckardt

Johann Wolfgang Goethe University Frankfurt am Main
D-60433 Frankfurt, Germany
roland@stuckardt.de

Abstract

In the last decade, much effort went into the design of robust third-person pronominal anaphor resolution algorithms. Typical approaches are reported to achieve an accuracy of 60-85%. Recent research addresses the question of how to deal with the remaining difficult-to-resolve anaphors. Lappin (2004) proposes a sequenced model of anaphor resolution according to which a cascade of processing modules employing knowledge and inferencing techniques of increasing complexity should be applied. The individual modules should only deal with and, hence, recognize the subset of anaphors for which they are competent.

It will be shown that the problem of focusing on the competence cases is equivalent to the problem of giving precision precedence over recall. Three systems for high precision robust knowledge-poor anaphor resolution will be designed and compared: a ruleset-based approach, a salience threshold approach, and a machine-learning-based approach. According to corpus-based evaluation, there is no unique best approach. Which approach scores highest depends upon type of pronominal anaphor as well as upon text genre.

1. Introduction

In the last decade, much effort went into the design of robust anaphor resolution algorithms. A considerable number of knowledge-poor approaches that operate on noisy data has been developed. According to the reported empirical evaluation results, these systems achieve an accuracy of 60-85% on third-person non-possessive and possessive pronouns.[1] Hence, still a considerable number of potentially difficult cases is not accounted for.

Taking this observation as the point of departure, recent research addresses the question of how to deal with these difficult-to-resolve instances of pronominal anaphora. Lappin (2004) proposes a *sequenced model of anaphor resolution* according to which a cascade of processing modules employing knowledge and inferencing techniques of increasing complexity should be applied: (1) the classical robust syntactic salience-based approach followed by (2) statistically measured lexical preference criteria and (3) abductive inferencing based on domain and world knowledge. At each stage of processing, all and only those anaphors should be assigned an antecedent for which the respective module is *competent*; anaphors that are beyond the horizon of the particular module should remain (locally) unresolved, i.e. left to the discretion of potentially more competent downstream modules. Thus, one question of central importance concerns the design of suitable *confidence measures* that support the recognition of these competence cases.

In the paper at hand, this important issue of *competence-oriented anaphor resolution* will be studied in detail. Recognizing and handling the competence cases only can be seen as equivalent to giving precision precedence over recall at the expense of (locally) leaving some of the anaphoric expressions unresolved. This problem has previously been addressed by the CogNIAC system of Baldwin (1997), which aims at high precision anaphor resolution. According to the research presented below, high precision strategies play a central role regarding the engineering of sequenced approaches to anaphor resolution. The study focuses on competence-oriented anaphor resolution through robust knowledge-poor approaches as typically employed upstream in sequenced architectures. It begins with an analysis of sequential vs. parallel anaphor resolution architectures, according to which biasing towards high precision plays an important role in conjunction with the sequential processing model. In the subsequent sections, three approaches to high precision robust knowledge-poor anaphor resolution will be specified, evaluated, and compared.

2. Anaphor Resolution Architectures

There are two basic ways of integrating different modules that solve, or contribute to, the solution of a particular natural language processing task: employing a *sequential* model, or employing a *parallel* model. The subsequent discussion focuses on the special case that all modules to be combined provide individual *complete* solutions of the natural language processing task to be performed. Concerning the task of anaphor resolution, this means that each of the system modules to be integrated performs *antecedent decisions* for a subset of anaphors to be resolved; the remaining anaphors, which can be considered to lie outside the competence space of the particular module, remain unresolved. Under this condition, precision and recall measures can be applied both at the level of these individual *competence modules* and at the level of the integrated system. The following analysis shows that the sequential model and the parallel model behave different with respect to the relation between individual and cumulated precision and recall.

2.1. Properties of the sequential model

Approaches following the sequential processing model employ a sequence of *competence modules* $CM_i, 1 \leq i \leq m$, each of which selects antecedents for a subset of anaphors with cardinality n_i, leaves the remaining anaphors

[1] Among relevant recent work are the manually designed approaches Lappin and Leass (1994), Kennedy and Boguraev (1996), Baldwin (1997), Mitkov (1998), Stuckardt (2001) and the machine-learning-based approaches Connolly et al. (1994), Aone and Bennett (1995), Ge et al. (1998), Soon et al. (2001), Stuckardt (2004)

unresolved and passes them on to downstream modules. Let N be the total number of anaphors to be interpreted, and c_i be the number of anaphors correctly resolved by CM_i. The sequential model possesses the important property that the *precision errors* $(n_i - c_i)$ of the individual competence modules sum up to the precision errors of the complete anaphor resolution module, i.e. once an anaphor is incorrectly resolved, there will be no review of this decision by any of the subsequent modules:

$$\sum_{i=1}^{m}(n_i - c_i) = \sum_{i=1}^{m} n_i - \sum_{i=1}^{m} c_i$$

This does *not* hold for the *recall errors*, since anaphors that are not locally handled by a certain module may be handled by any of the subsequent modules, thus reducing the recall errors at the global level:

$$\sum_{i=1}^{m}(N - \sum_{j=1}^{i-1} n_j - c_i) =$$

$$= N - \sum_{i=1}^{m} c_i + (m-1)N - \sum_{i=1}^{m-1}\sum_{j=1}^{i} n_j \geq$$

$$\geq N - \sum_{i=1}^{m} c_i$$

(The inequality holds because each of the sums $\sum_{j=1}^{i} n_j, 1 \leq i \leq m-1$ (i.e., the number of anaphors processed prior to competence module i) is lower than or equal to the total amount of anaphors N.) Hence, in the sequential model, biasing the individual competence modules towards high precision crucially contributes to the design of an anaphor resolution system that, at global level, may achieve high precision *and* recall even if, at local level, the individual recall may be reduced.

2.2. Properties of the parallel model

According to the parallel processing model, there is no priority ordering of the modules. Hence, for a particular anaphor α, more then one competence module may make an antecedent prediction, and these predictions may contradict each other. Employing an appropriate evidence combination scheme amounts to a form of competence mediation between the different modules that are integrated by the parallel model.
Due to this underspecification, there is no straightforward relation between precision and recall of the individual competence modules (which can be computed as above) and cumulated precision and recall of the anaphor resolution system.

2.3. Implications

From a point of view of natural language engineering, the sequential model offers advantages since substantial efforts concerning the design, optimization, and implementation of suitable competence integration techniques are circumvented. Concerning anaphor resolution, Lappin (2004) points out specific advantages that are to be expected by *sequentially* combining, in the order of mention, a robust syntactic salience-based competence module with a module employing statistical lexical preference evidence and a module that performs enhanced abductive inferences based on domain and world knowledge. According to the above considerations, achievement of this goal is facilitated by appropriately biasing the syntactic salience-based competence module towards *high precision*, trying to leave the anaphors outside its scope of competence unresolved in order to pass them on to the downstream competence modules. In the following sections, this particular problem of competence-oriented anaphor resolution will be investigated in detail.

3. Three Approaches to Robust High Precision Anaphor Resolution

The robust syntactic salience-based anaphor resolution system ROSANA and its machine-leaning-based descendant ROSANA-ML are taken as the starting points (see (Stuckardt, 2001; Stuckardt, 2004)). Based on these implementations, three knowledge-poor high precision anaphor resolution systems are designed which work robustly on potentially noisy data.

3.1. ROSANA-CogNIAC

The first and most obvious approach consists in the reimplementation of the CogNIAC system of Baldwin (1997), which is specifically designed to achieve high precision anaphor resolution. CogNIAC combines the morphological agreement and syntactic disjoint reference filters with a set of six manually crafted antecedent selection rules, each of which covering one particular configuration in which there seems to be little or no ambiguity regarding the available antecedent candidates. The rules are applied in order of increasing ambiguity: if a rule fires (i.e. applies), the respective candidate will be chosen; if none of the rules applies, the anaphor remains unresolved.
According to the CogNIAC strategy, the text is processed strictly from left to right, taking into account only candidates that precede the anaphor under consideration, i.e. candidates from the read-in portion of the discourse. The antecedent filters are applied prior to the six high precision rules, which are:

(CR1) *unique in discourse*: if only a single candidate from the read-in portion of the discourse remains, then choose this candidate;

(CR2) *reflexive pronouns*: if the anaphor is a reflexive pronoun, then choose the nearest possible antecedent in the read-in portion of the current sentence;

(CR3) *unique in current and prior*: if, after application of the antecedent filters, only a single candidate from the prior and the read-in portion of the current sentences remains, then choose this candidate;

(CR4) *possessive pronouns*: if the anaphor is a possessive pronoun and there is a single exact string match of the possessive in the prior sentence, then pick this match as the antecedent;

(CR5) *unique in current*: if there is a single possible antecedent in the read-in portion of the current sentence, then pick it as the antecedent;

(CR6) *unique subject/subject pronoun*: if the anaphor is the subject of the current sentence, and the subject of the prior sentence contains a single possible antecedent, then pick it as the antecedent.

The CogNIAC rules and processing strategy have been reimplemented, using the anaphor resolution system ROSANA as the point of departure (see (Stuckardt, 2001)). The robust operationalizations of the syntactic disjoint reference and morphological agreement filters are adopted from ROSANA; the salience-based antecedent selection strategy is substituted by the CogNIAC rules, which are sequentially applied to each of the anaphors to be resolved; anaphors are considered from left to right.

3.2. Employing a salience threshold: ROSANA-θ

A second approach to achieve high precision anaphor resolution (which, in fact, can be considered as the *baseline strategy*) consists in an even more immediate adaptation of the antecedent selection phase of classical, salience-based anaphor resolution algorithms:

> *Given a salience threshold θ, only such candidates are taken into account the salience of which exceeds the threshold θ.*

The rationale behind this heuristics is that salience does not only form a base for heuristically comparing the relative plausibility of two candidates (and choosing the one with higher salience); in addition, it can be employed non-relationally as an heuristic estimate of the probability that an individual candidate is a correct antecedent, thus allowing to decline candidates with low salience in order to avoid risky decisions.

Again, ROSANA is taken as the point of departure. By modifying its antecedent selection step in a straightforward way, the system ROSANA-θ is obtained that employs the above strategy. The actual degree of precision biasing depends upon the value of θ, thus leaving room for different tradeoffs.

3.3. ROSANA-ML tuned towards high precision

Another way to achieve high precision anaphor resolution has been investigated during the empirical experiments with the machine-learning-based approach ROSANA-ML, which employs decision tree classifiers for selecting among antecedent candidates fulfilling the filtering criteria (see (Stuckardt, 2004)). Basically, the decision trees represent classifier functions which map pairs of anaphors and antecedent candidates (represented as feature vectors) to a prediction $\in \{COSPEC, NON\_COSPEC\}$. Besides the primary classification result, the leaves of the C4.5 decision trees contain additional quantitative information. Specifically, each leaf provides the total number μ of *training cases* that match the respective decision path, and the number $\varepsilon \leq \mu$ of these cases that are, through the category prediction of the leaf, *wrongly classified*. By computing the quotient $\frac{\varepsilon}{\mu}$, it should thus be possible to derive an estimate

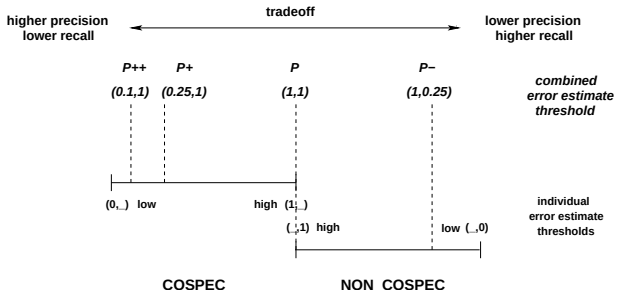

Figure 1: ROSANA-ML-θ: error estimate threshold

of the classification error probability of the particular leaf. The base version of the ROSANA-ML algorithm prefers candidates predicted to COSPECify over candidates predicted to NON_COSPECify,[2] and employs surface-topological distance as the secondary criterion (see (Stuckardt, 2004)). By looking at the quotient $\frac{\varepsilon}{\mu}$, this preference criterion can be refined as follows: prefer COSPEC candidates over NON_COSPEC candidates; at the secondary level, prefer COSPEC candidates with smaller classification error estimate $\frac{\varepsilon}{\mu}$ over COSPEC candidates with higher $\frac{\varepsilon}{\mu}$, and prefer NON_COSPEC candidates with higher classification error estimate $\frac{\varepsilon}{\mu}$ over NON_COSPEC candidates with lower $\frac{\varepsilon}{\mu}$. Finally, by setting a *threshold* $\theta := (\theta_{co}, \theta_{nonco})$, as illustrated in figure 1, i.e. by eliminating all COSPEC candidates the classification error estimate of which falls *above* θ_{co}, and by eliminating all NON_COSPEC candidates the classification error estimate of which falls *below* θ_{nonco}, a bias can be imposed that gradually trades off recall for precision.

4. Text Corpus and Formal Evaluation

The evaluation runs will be performed on a corpus of 66 news agency press releases, comprising 24,712 words, 406 third-person non-possessives, 246 third-person possessives, 172 relative pronouns, and 13 reflexive pronouns. The corpus is partitioned into a training subset (31 documents, 11,808 words, 202 non-possessives, 115 possessives, 71 relative pronouns, 9 reflexives) and an evaluation subset (35 documents, 12,904 words, 204 non-possessives, 131 possessives, 101 relative pronouns, 4 reflexives). All experiments take place without manual intervention, i.e. without a-priori correction of noisy input, which may contain orthographic, syntactic, or preprocessing/parsing errors.

The anaphor resolution performance will be evaluated in the evaluation discipline of *immediate antecedency*. Let (P, A) be a pair consisting of a pronominal occurrence P and the antecedent occurrence A determined by the anaphor resolution system. Let o_{++} be the set of pairs where P and A belong to the same key equivalence class, and o_{+-} the set of pairs that, according to the key, do not cospecify. Furthermore, let $o_{+\_}$ be the set of cases where no antecedent has been determined (A is empty), and $o_{+?}$ be the set of instances where A does not correspond to a valid key occurrence. By drawing the usual distinction between precision

[2] Since the decision tree classifiers employed by ROSANA-ML are designed to substitute the antecedent *preference* criteria only, their predictions are not used to further *filter* the candidate set.

and recall, according to which, in the latter case, one has to take into account empty antecedents A as well, one obtains the following definitions:[3]

$$P := \frac{|o_{++}|}{|o_{++}| + |o_{+-}| + |o_{+?}|},$$

$$R := \frac{|o_{++}|}{|o_{++}| + |o_{+-}| + |o_{+?}| + |o_{+\_}|}$$

5. Experiments and Empirical Results

In the table shown in figure 2, the results of the formal, corpus-based evaluation on the *News Agency Press Releases* corpus are summarized. Precision and recall figures for four different types of third-person pronoun anaphora - nonpossessives (PER3), possessives (POS3), relative pronouns (RELA), and reflexives (REFL) - are displayed. Results for relative pronouns and reflexives are given because they are dealt with, too, by the CogNIAC resolver of Baldwin (1997).[4] Since, however, the antecedent choice for these pronouns is tightly governed by easy-to-implement syntactic constraints, evaluation results are in general better than for nonpossessives and possessives. Thus, these latter two types of pronouns constitute the anaphor resolution test cases proper, and the subsequent discussion will focus on them.

The results of the salience-based robust anaphor resolution system ROSANA are displayed for reference purposes (experiment (0)). ROSANA attempts to assign to every pronoun an antecedent. As a consequence, for non-possessives as well as for possessives, precision equals recall (0.71 and, respectively, 0.76).[5]

5.1. ROSANA-CogNIAC

As described above, the CogNIAC rules (CR1) to (CR6) have been implemented as the antecedent selection device in an anaphor resolution system which employs the robust antecedent filter implementation (agreement, syntactic disjoint reference) of ROSANA. As indicated by the results of a series of early experiments on the training data portion of the corpus, some care should be exercised in order to arrive at a valid implementation of the CogNIAC rules. First and most importantly, it has to be taken into account that, while the employed notions of anaphor, candidate, and antecedent partly refer to the *occurrence* level (e.g., (CR4) employs surface string identity, and (CR6) refers to the syntactic function of particular occurrences), the uniqueness conditions that govern some of the rules apply at *discourse referent (i.e. coreference class)* level. Second, it turned out that, in requiring a "*single exact string match of the possessive in the prior sentence*", viz. string identity at the level of the possessee, the governing condition of rule (CR4), which is satisfied in cases like the following

> *After he was dry, Joe carefully laid out the damp towel in front of* **his locker**. *Travis went over to* **his locker**, *took out a towel and started to dry off.*

(example taken from (Baldwin, 1997), p. 40) is too strong. Since the original version of (CR4) fired only for one of the 115 instances of possessive pronouns occurring in the *News Agency Press Releases* training corpus, it was relaxed by restricting the string match condition to apply merely to the possessive pronoun itself, thus yielding rule (CR4)'.

According to the results of experiment *(1) ROSANA-CogNIAC, (CR4)'* given in figure 2, results are unsatisfying for the most frequent pronoun type of non-possessive pronouns as the gap between 0.66 precision at 0.49 recall and the CogNIAC average of 0.92 precision at 0.64 recall (according to Baldwin (1997)) is still considerable. In order to see what goes wrong, an in-depth qualitative analysis of the 62 instances of wrongly resolved third-person pronouns was carried out. The analysis revealed that only 10 of these cases are due to the heuristical nature of the CogNIAC rule set, e.g.[6]

> *It was not until 1992 that the discovery of lost documents forced* <u>*the government*</u> *to acknowledge that* **the Imperial Army** *made the women serve as sex slaves. Prior to that,* **it** *had acknowledged the existence of wartime brothels but said women worked willingly as prostitutes.*

In this example, rule (CR6) led to the selection of a wrong antecedent (*the Imperial Army*), which occupies a subject role in the prior sentence; the non-subject candidate *the government*, which would have been correct here, is not considered by any of the CogNIAC rules.

Another 13 of the wrongly resolved cases can be attributed to incorrect (heuristically assigned) gender information which led to non-firing and/or wrongly firing CogNIAC rules. 4 cases are attributable to errors at the stage of robust preprocessing (occurrence identification, parsing errors, etc.). Importantly, another class of 15 cases was identified that can be accounted for by a modified version of (CR6). According to the original rule statement, (CR6) only looks for candidates in the subject role of the *prior* sentence. However, there are numerous cases in which this prediction yields wrong results; cases like the following illustrate that this can be partially accounted for by looking at subject candidates of the *current* sentence, too:[7]

[3] If *all* pronouns are resolved, set $o_{+\_}$ is empty. Under this condition, precision equals recall, and these measures yield results that are identical with the accuracy figures given in the evaluations of the classical approaches of, e.g., Lappin and Leass (1994) and Kennedy and Boguraev (1996).

[4] Since ROSANA-ML-θ only resolves nonpossessives and possessives by machine learning means, no evaluation figures are given for relative pronouns and reflexives. Essentially, they are identical with the figures determined for ROSANA (salience-based).

[5] For relative pronouns and reflexives, nevertheless, recall is to some extent lower than precision, which is a direct consequence of the tight syntactic conditions that govern the antecedent choices of these two types of pronouns; in conjunction with preprocessing (particularly parsing) problems, this might result in cases in which pronouns remain unresolved as there seem to be no admissible candidates available.

[6] In the following examples of wrongly resolved anaphors, the pronoun and the selected (incorrect) antecedent are displayed in bold face, and the antecedent that would have been correct is underlined.

[7] A comparative analysis revealed that many of the cases in which there is a local subject antecedent are instances of *indirect*

experiment	antecedents (P, R), *News Agency Press Releases* corpus			
	PER3	POS3	RELA	REFL
(0) ROSANA (salience-based)	$(0.71, 0.71)$	$(0.76, 0.76)$	$(0.78, 0.73)$	$(1.00, 0.75)$
(1) ROSANA-CogNIAC, (CR4)'	$(0.66, 0.49)$	$(0.82, 0.53)$	$(0.84, 0.68)$	$(1.00, 0.75)$
(2) ROSANA-CogNIAC, (CR4)', (CR6)'	$(0.74, 0.59)$	$(0.82, 0.53)$	$(0.83, 0.70)$	$(1.00, 0.50)$
(3) ROSANA-θ (θ = 90)	$(0.75, 0.67)$	$(0.79, 0.74)$	$(0.84, 0.73)$	$(1.00, 0.75)$
(4) ROSANA-θ (θ = 110)	$(0.79, 0.62)$	$(0.81, 0.50)$	$(0.84, 0.73)$	$(1.00, 0.75)$
(5) ROSANA-ML-θ, p (θ = (1.00,1.00))	$(0.79, 0.51)$	$(0.86, 0.60)$	*as (0) ROSANA*	*as (0) ROSANA*
(6) ROSANA-ML-θ, p^- (θ = (1.00,0.25))	$(0.74, 0.56)$	$(0.78, 0.63)$		
(7) ROSANA-ML-θ, p^+ (θ = (0.25,1.00))	$(0.81, 0.45)$	$(0.89, 0.50)$		
(8) ROSANA-ML-θ, p^{++} (θ = (0.10,1.00))	$(0.83, 0.31)$	$(1.00, 0.17)$		

Figure 2: evaluation results, *News Agency Press Releases* corpus

> *Officials of the rival Korean states did agree to hold a third round of talks in October and to continue their efforts to reach agreement on providing rice assistance to the North,* **which** *is apparently suffering from food shortages. South Korean sources said the talks ran into some difficulties because* <u>*Seoul*</u> *said* **it** *wanted to extend the agenda beyond the issue of rice aid to include economic cooperation.*

(CR6) has been correspondingly modified, yielding rule (CR6)'. In conjunction with a tie breaking heuristics according to which, if there is more than one subject antecedent candidate, the topologically nearest is preferred, the performance on third-person non-possessives is considerably enhanced (see row *(2) ROSANA-CogNIAC, (CR4)', (CR6)'* of table 2), amounting to 0.74 precision at 0.59 recall.

In figure 3, the individual performance figures of rules (CR1) to (CR6)' (rules (CR4) and (CR6) modified as described) are given. The entries *x/y* of the table indicate how many pronouns of a certain type a particular rule attempted to resolve (*y*), and how many of these antecedent decisions were correct (*x*); the figure in brackets denotes the respective *precision* rate. According to the general strategy of applying the rules in the order (CR1) to (CR6)', the rows read from left to right, i.e. only pronouns not handled by a particular rule are passed on to the next rule to the right. In the last column, the number of anaphors not resolved by any of the rules is given. Concerning relative pronouns, only those 4 instances are included which are resolved through one of the high precision rules; the vast majority of cases (97 relatives) is resolved outside the high precision ruleset by purely configurational means and hence not shown in the table.

In the final row of the table, the cumulated performance figures of the different rules are given. Individual precision varies between 1.0 (rule (CR2) for reflexives) and 0.69 (rule (CR6)'). These results, which average to a total of 0.76 precision at 0.54 recall, are considerably worse than the results given in the original work of Baldwin (1997), in which a cumulated precision of 0.92 at 0.64 recall was determined. If one takes into account the 97 instances of relative pronouns that are not counted in figure 3 as they are resolved outside the high precision ruleset by purely configurational means, the performance raises to 0.78 precision at 0.60 recall, which is still worse than the original CogNIAC performance. The reduced performance can be partially attributed to the harder conditions under which ROSANA-CogNIAC was run as there have been no manual corrections of preprocessing results as Baldwin (1997) did in order to allow for direct comparison with the algorithm of Hobbs (1978). Moreover, the evaluation corpus genre (press releases) considerably differs from the genre of Baldwin's evaluation corpus (narrative texts of two persons of the same gender). In section 5.4., this latter issue will be further investigated.

5.2. ROSANA-θ

According to the experimental results *(3) ROSANA-θ (θ = 90)* and *(4) ROSANA-θ (θ = 110)* in figure 2, employing a threshold for biasing the salience-based ROSANA approach towards precision yields reasonable results. Setting the salience threshold θ to 90 brings a gain of 4%/3% precision at the expense of 4%/2% recall (for non-possessives/possessives, respectively); setting it to 110 brings 8%/5% precision at the expense of 9%/26% recall. In accordance with expectations, this gives evidence that the probability that a particular antecedent candidate is correct correlates with the salience degree of the respective discourse referent.

5.3. ROSANA-ML-θ

In the lower four rows of the table in figure 2, the results of a series of experiments with different threshold settings regarding the decision tree classification error estimates of ROSANA-ML are displayed. The four experiments (labeled p^-, p, p^+, p^{++}) differ with respect to the degree to which the threshold settings aim at trading off recall for precision, where p^- stands for lowest extent, and p^{++} stands for highest extent.

The basic precision bias setting of the experiment labeled p allows $\frac{\varepsilon}{\mu}$ values of ≤ 1 for COSPEC-predicted instances, and > 1 for NON_COSPEC-predicted instances; in other words, all and only those candidates are eliminated that are predicted not to cospecify (see figure 1). The precision bias can be weakened by eliminating only those candidates the NON_COSPEC prediction of which is incorrect with estimated probability falling below a threshold $\theta < 1$,

speech. It it possible that the original CogNIAC system handles these cases seperately, i.e. independently of the high precision ruleset (see (Baldwin, 1997)). If an independent module for interpreting indirect speech is available, (CR6) can possibly remain unchanged.

type	(CR1)		(CR2)		(CR3)		(CR4)'		(CR5)		(CR6)'		unres.
PER3	2/3	(0.66)	0/0	(-)	28/34	(0.82)	0/0	(-)	37/50	(0.74)	53/76	(0.70)	41
POS3	4/5	(0.8)	0/0	(-)	26/32	(0.81)	9/11	(0.82)	30/36	(0.83)	0/0	(-)	47
RELA	0/0	(-)	0/0	(-)	0/0	(-)	0/0	(-)	0/0	(-)	2/4	(0.5)	15
REFL	1/1	(1.0)	1/1	(1.0)	0/0	(-)	0/0	(-)	0/0	(-)	0/0	(-)	2
	7/9	(0.78)	1/1	(1.0)	54/66	(0.82)	9/11	(0.82)	67/86	(0.78)	55/80	(0.69)	105

Figure 3: ROSANA-CogNIAC rules, application results, *News Agency Press Releases* corpus

experiment	antecedents (P, R), *Mozart Operas* corpus			
	PER3	POS3	RELA	REFL
(A) ROSANA (salience-based)	$(0.79, 0.79)$	$(0.77, 0.77)$	$(0.95, 0.90)$	$(0.83, 0.83)$
(B) ROSANA-CogNIAC, (CR4)'	$(0.85, 0.61)$	$(0.83, 0.58)$	$(1.00, 0.76)$	$(1.00, 1.00)$
(C) ROSANA-CogNIAC, (CR4)', (CR6)'	$(0.81, 0.61)$	$(0.84, 0.59)$	$(0.94, 0.76)$	$(1.00, 1.00)$
(D) ROSANA-CogNIAC, (CR6)'	$(0.81, 0.61)$	$(0.92, 0.55)$	$(0.94, 0.76)$	$(1.00, 1.00)$
(E) ROSANA-θ (θ = 90)	$(0.80, 0.69)$	$(0.80, 0.77)$	$(1.00, 0.86)$	$(1.00, 1.00)$
(F) ROSANA-θ (θ = 110)	$(0.83, 0.53)$	$(0.76, 0.36)$	$(1.00, 0.81)$	$(1.00, 1.00)$

Figure 4: evaluation results, *Mozart Operas* corpus

e.g. $\theta = 0.25$ (experiment p^-). Similarly, the bias can be strenghtened by imposing lower error ratio thresholds for the candidates predicted to COSPECify and eliminating all candidates predicted to NON_COSPECify (experiments p^+ and p^{++}). In accordance with expectations, the scores for the different settings indicate that, by employing the quantitative information given at the decision tree leaves in the above-described way, one obtains a suitable (albeit heuristical) means for gradually biasing ROSANA-ML towards high precision.

5.4. Comparison

An inter-method comparison of the results displayed in figure 2 leads to a non-uniform assessment. Regarding *(2) ROSANA-CogNIAC, (CR4)', (CR6)'* vs. *(3) ROSANA-θ (θ = 90)*, the latter approach yields better (P, R) tradeoffs for nonpossessives; for possessive pronouns, however, none of the tradeoffs majorizes the other. If the higher threshold is chosen (*(4) ROSANA-θ (θ = 110)*), the former approach slightly outperfoms the latter on possessives. Hence, the evaluation results indicate that neither of these approaches is generally superior to the other. According to the results that have been determined for *ROSANA-ML-θ*, this machine-learning-based approach seems to yield the by far best tradeoffs for possessives (p setting); regarding nonpossessives, however, it seems to be outperformed by *ROSANA-θ*. Hence, based on the available experimental results, it is by now impossible to single out a unique best approach.

Additional evidence can be gained by looking at another evaluation corpus of a different text genre. Figure 4 displays the results of the high precision methods ROSANA-CogNIAC and ROSANA-θ on a corpus of three texts describing the plots of Mozart operas.[8] Most importantly, original CogNIAC's rule (CR6) proved to be superior to the modified rule (CR6)'. If, again, we compare the approaches ROSANA-CogNIAC and ROSANA-θ, we observe that, on nonpossessives, the latter is no longer superior to the former. Another experiment showed that employing the original (strong) version of the possessive rule (CR4) can be considered reasonable when processing texts of the *Operas* genre.

It is further instructive to compare the performance of the three approaches at the level of particular anaphors. In determining whether there is a distinguished subclass of anaphors the instances of which are wrongly (or correctly) resolved regardless of the employed approach, one gets a better understanding regarding the relative amount of *inherently difficult* (or *inherently easy*) cases to be expected. Similarly, by analyzing in how much the individual predictions differ, one gains essential information concerning the individual strengths and weaknesses of the approaches. If, for reasons of expository simplicity, one restricts the analysis to the *pairwise* comparison of the three approaches, then the investigation is based on *contingency tables* which display the cardinalities of the intersection sets of (1) correctly resolved anaphors (++), (2) wrongly resolved anaphors (+-, +?), and (3) unresolved anaphors (+_) (see section 4.).

In figure 5, three two-dimensional contingency tables are given that have been computed by pairwise comparing three of the above discussed high precision approaches: *(2) ROSANA-CogNIAC, (CR4)', (CR6)', (3) ROSANA-θ (θ = 90), and (5) ROSANA-ML-θ, p*. The tables show the cumulated figures for the third person nonpossessive and possessive pronouns that occur in the evaluation part of the *News Agency Press Releases* corpus.[9] The rows correspond to the approach that is mentioned first in the head of the table; the columns correspond to the second-mentioned approach. According to the figures in the main diagonals of

[8] Since this corpus is too small, it proved to be impossible to evaluate the machine-learning-based approach ROSANA-ML-θ on it, which requires a reasonable amount of training data.

[9] The evaluation part of the *News Agency Press Releases* corpus contains 204 nonpossessives and 131 possessives. In each of the three matrices, only those pronouns are taken into account that are "scoreably interpreted" by *both* of the respective two systems: an unresolved pronoun is not counted if the respective coreference tag in the key is marked as optional; hence the small differences regarding the total number of nonpossessives/possessives in the three tables.

R-CogN. (CR4,6)' vs. R-θ (90)				
	++	+-/?	+_	n
++	180	3	6	189
+-/?	13	39	5	57
+_	41	29	18	88
n	234	71	29	334

R-CogN. (CR4,6)' vs. R-ML-θ p				
	++	+-/?	+_	n
++	134	6	49	189
+-/?	18	24	14	56
+_	29	11	48	88
n	181	41	111	333

R-ML-θ p vs. R-θ (90)				
	++	+-/?	+_	n
++	158	11	12	181
+-/?	7	32	2	41
+_	69	27	15	111
n	234	70	29	333

Figure 5: contingency tables, PER3/POS3 pronouns, *News Agency Press Releases* corpus

the contingency tables, any two of the high precision approaches yield equivalent predictions for more than 50% of the individual anaphoric instances.[10] However, the agreement between *(2) ROSANA-CogNIAC, (CR4)', (CR6)'* and *(3) ROSANA-θ (θ = 90)* is considerably higher than in the two other comparisons in which *(5) ROSANA-ML-θ, p* is involved. More specifically, results illustrate that the former two systems perform correctly on a large common set of anaphors. The figures in the main diagonals indicate that there is a core set of about 100 to 130 easy-to-resolve anaphors, and there may be another quite small core set of difficult-to-resolve anaphors (clearly below 80, as can be seen by looking at the number of cases in the $[+-/?, +\_]$ submatrix of the rightmost contingency table). According to the further figures of the leftmost contingency table, *(3) ROSANA-θ (θ = 90)* can be seen as more or less majorizing the approach *(2) ROSANA-CogNIAC, (CR4)', (CR6)'*.

6. Conclusion and Further Research

The evaluation results provide first evidence that the performance of the CogNIAC ruleset highly depends on the genre of the texts to be referentially interpreted. A comparative evaluation indicates that the two other competence-oriented approaches to robust knowledge-poor anaphor resolution achieve reasonable (P, R) tradeoffs as well. Which approach actually performs best seems to vary with the type of pronominal anaphor to be resolved as well as across text genres.

However, an informed choice of the most suitable upstream (P, R) tradeoff should not only be based on quantitative considerations. It has been further shown that contingency analysis is a powerful method for developing a proper understanding of the relative performance of competence modules. The important issue of optimally combining competence modules according to the sequential processing model should be subjected to further research. Moreover, the question which high precision approach performs best for which text genre and for which type of anaphoric expression should receive further attention.

7. References

Chinatsu Aone and Scott William Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the ACL, Santa Cruz, New Mexico*, pages 122–129.

Breck Baldwin. 1997. Cogniac: High precision coreference with limited knowledge and linguistic resources. In Ruslan Mitkov and Branimir Boguraev, editors, *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphor Resolution for Unrestricted Texts, Madrid*, pages 38–45.

Dennis Connolly, John D. Burger, and David S. Day. 1994. A machine-learning approach to anaphoric reference. In *Proceedings of the International Conference on New Methods in Language Processing (NEMLAP)*.

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

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–338.

Christopher Kennedy and Branimir Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pages 113–118.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Shalom Lappin. 2004. A sequenced model of anaphora and ellipsis resolution. In António Branco, Tony McEnery, and Ruslan Mitkov, editors, *Anaphora Processing: Linguistic, Cognitive, and Computational Modelling*, page (forthcoming). John Benjamins.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING'98/ACL'98), Montreal*, pages 869–875.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Roland Stuckardt. 2001. Design and enhanced evaluation of a robust anaphor resolution algorithm. *Computational Linguistics*, 27(4):479–506.

Roland Stuckardt. 2004. A machine learning approach to preference strategies for anaphor resolution. In António Branco, Tony McEnery, and Ruslan Mitkov, editors, *Anaphora Processing: Linguistic, Cognitive, and Computational Modelling*, page (forthcoming). John Benjamins.

[10] *Evaluation result agreement* is defined as follows: if system S_A selects a cospecifying antecedent then system S_B does, too (this requires identity at discourse referent level rather than antecedent occurrence level); if system S_A selects a non-cospecifying antecedent, then also system S_B selects a non-cospecifying antecedent (which may refer to an arbitrary discourse referent); if system S_A doesn't select an antecedent, then neither S_B does.