

A MACHINE LEARNING APPROACH TO PREFERENCE STRATEGIES FOR ANAPHOR RESOLUTION

ROLAND STUCKARDT

Johann Wolfgang Goethe University Frankfurt am Main¹

Abstract

In the last few years, much effort went into the design of robust anaphor resolution algorithms. Many algorithms are based on antecedent filtering and preference strategies that are manually designed. Along a different line of research, corpus-based approaches have been investigated that employ machine learning or statistical techniques for deriving strategies automatically, thus considerably facilitating knowledge engineering. Since, however, manually designing the robust antecedent filtering strategies constitutes a once-for-all effort, the question arises whether at all they should be derived automatically.

In this article, it is investigated what may be gained by combining the best of two worlds: designing the universally valid antecedent filtering strategies manually, and deriving the potentially genre-specific antecedent preference strategies automatically by applying machine learning techniques. Following this paradigm, an anaphor resolution system ROSANA-ML is designed. Through a thorough formal evaluation, it is shown that, while exhibiting additional advantages, ROSANA-ML performs similar to its manually designed ancestor ROSANA.

1. Introduction

The interpretation of textual anaphoric expressions is a subtask which is crucial to a wide range of natural language processing problems. In the last few years,

¹ Im Mellsig 25, D-60433 Frankfurt am Main, Germany
E-mail: roland@stuckardt.de. Phone: +49 (0)69 517797

much effort went into the design of robust, knowledge-poor algorithms that are capable of processing potentially noisy data. Many approaches take as a starting point the landmark work of Lappin & Leass (1994), in which an algorithm for interpreting third person pronouns is developed that relies upon the idealistic assumption that, for the sentences to be interpreted, complete syntactic parses are available. For achieving robustness, various solutions have been suggested, e.g. to employ a robust part-of-speech tagger instead of full syntactic parsing (Kennedy & Boguraev 1996), or to generalize the strategies to work on possibly fragmentary syntactic descriptions (Stuckardt 2001; Stuckardt 1997).

Along a different line of research, corpus-based approaches have been investigated that employ machine learning or statistical techniques for deriving anaphor resolution strategies automatically (Soon, Ng & Lim 2001; Paul, Yamamoto & Sumita 1999; Ge, Hale & Charniak 1998; Aone & Bennett 1996; Aone & Bennett 1995; Dagan, Justenson, Lappin, Leass & Ribak 1995; McCarthy & Lehnert 1995; Connolly, Burger & Day 1994). These approaches are considered particularly attractive because the effort for designing and implementing the strategies is reduced. In general, the automatic derivation of anaphor resolution strategies relies upon the availability of sufficiently large text corpora that are tagged, in particular, with referential information.²

E.g., Aone & Bennett (1995) employ supervised decision tree learning for deriving an anaphor resolution master strategy that covers antecedent filtering as well as antecedent preference criteria. They primarily aim at providing an elegant solution to the robustness issue per se; as an important advantage, they point out that their approach automatically generalizes to additional types of anaphoric expressions. However, the inventory of relevant types of anaphoric expressions is limited. Moreover, recent research has revealed that some classical approaches to robust anaphor resolution which descend from the work of Lappin & Leass (1994) are, with respect to the robust operationalization of the antecedent filtering strategies of syntactic disjoint reference and agreement in person/number/gender, nearly optimal (Stuckardt 2001; Kennedy & Boguraev 1996). Since the robust implementation of these successful anaphor resolution strategies constitutes a once-for-all effort, the question arises whether at all they should be derived automatically through the application of machine learning techniques.

² While some approaches employing unsupervised learning have been explored, the most promising ones make use of supervised techniques. Some referentially annotated corpora have been developed during the last few years (particularly for the DARPA Message Understanding Conferences (MUCs)). However, the total amount of available tagged texts is still quite restricted.

A MACHINE LEARNING APPROACH TO PREFERENCE STRATEGIES

In the present article, it is investigated what may be gained by employing machine-learned *preference* strategies³ as part of a robust anaphor resolution approach according to the Lappin & Leass (1994) paradigm in which the antecedent filtering strategies are manually designed. The algorithm ROSANA described in (Stuckardt 2001) is taken as the starting point. Empirical studies in this article have shown that, for achieving optimal interpretation results, the antecedent preference strategies, which come as sets of *weighted salience factors*, should be designed genre-specifically, since text genres seem to differ with respect to the characteristic properties of their typical coherence structures. Hence, there is no once-for-all optimal design of preference heuristics. Consequently, antecedent preference strategies are ideal targets for applying machine learning techniques.

Thus, it is explored what may be gained by combining the best of two worlds: designing the universally valid antecedent filtering strategies manually - once and for all -, and deriving the genre-specific antecedent preference strategies automatically by applying machine learning techniques. An anaphor resolution system ROSANA-ML, which follows this paradigm, is designed and implemented. Through a thorough formal evaluation, it is shown that, with respect to two important evaluation measures, ROSANA-ML reaches a level of performance that compares with the interpretation quality of its manually designed ancestor ROSANA. More specifically, the evaluation reveals that, whereas regarding third person possessive pronouns, a gain is achieved, the results regarding third person non-possessives slightly lag behind the performance of the manually designed system. In particular, the evaluation results regarding non-possessives indicate that the set of features over which the classifiers are learned should be suitably supplemented; it is expected that this enhances the need for still larger corpora of referentially annotated training texts, thus confirming similar findings of other researchers (e.g. (Mitkov 2001)). Moreover, the results of a series of further experiments indicate that, regarding third-person pronominal anaphora in English, by biasing ROSANA-ML towards precision, better (precision, recall) tradeoffs (henceforth referred to as (P,R) tradeoffs) can be obtained than those determined by Aone & Bennett (1995) for the case of Japanese zero pronouns.

³ In emphasizing the application case of these criteria during anaphor resolution (viz., the antecedent selection phase, see section 3.2), one could equally well speak of antecedent *selection* strategies.

The article is organized as follows. In section 2, the fundamental methodology is described. In particular, the machine learning approach, which employs the C4.5 decision tree algorithm of Quinlan (1993), is outlined; moreover, it is sketched how the training data are obtained and how the learned decision trees are applied for selecting antecedents. In section 3, formal specifications of the algorithms are given, and the underlying paradigm of learning preference strategies is further illustrated; an implementation, the ROSANA-ML system, is briefly described. In section 4, a series of experiments regarding, in particular, the choice of features, the employed training strategies, and the learning performance is designed. In section 5, the respective empirical evaluation results are interpreted. Finally, in sections 6 and 7, the findings are compared with the results of other approaches to anaphor resolution, and promising directions of further research are identified.

2. Methodology

In Figure 1, the machine learning approach to anaphor resolution followed by ROSANA-ML is outlined. It is distinguished between the training phase, which is shown in the upper part of the figure, and the application (anaphor resolution) phase sketched in the lower part of the figure.

During the *training phase*, based on a training text corpus, a set of feature vectors is generated which consists of feature tuples derived from the (*anaphor*, *antecedent candidate*) pairs that are considered during the antecedent selection phase of the anaphor resolution algorithm ROSANA. This output is written to a file *data.fve*, which, during the next step, is classified by employing intellectually gathered key data (file *data.key*). The result consists of a set of training vectors (file *data.fvc*) which are classified as either COSPEC or NON_COSPEC, depending on whether, according to the key, the respective occurrences of anaphor and antecedent candidate are *cospecifying* or *not cospecifying*. Finally, these training cases are submitted to the C4.5 machine learning algorithm: C4.5 derives a decision-tree-shaped classifier (file *data.dts*) suitable for categorizing arbitrary feature vectors that are of the same signature as the training vectors.

In the *application (anaphor resolution) phase*, the learned classifiers are employed for antecedent selection: to discern between more and less plausible candidates, instead of applying a set of salience factors (as done by the manually designed algorithm ROSANA), a decision tree lookup is performed, which yields a (heuristic) prediction COSPEC or NON_COSPEC. In combination with a

A MACHINE LEARNING APPROACH TO PREFERENCE STRATEGIES

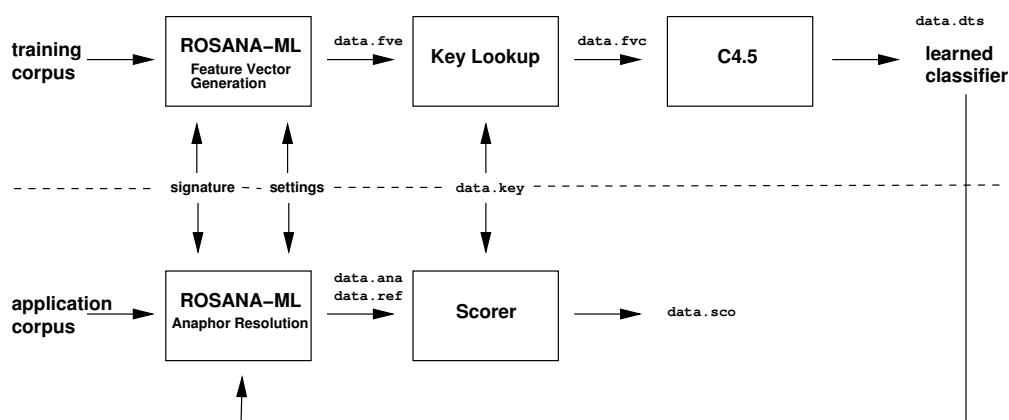


Figure 1: ROSANA-ML: training phase vs. application (anaphor resolution) phase

secondary preference criterion (such as surface distance), this prediction renders possible an ordering of the antecedent candidates of an anaphor according to decreasing plausibility. The anaphor resolution output is written to the files *data.ana* (coreference classes) and *data.ref* (basically, anaphoric resumption chains). During *formal evaluation*, the interpretation quality of ROSANA-ML will be measured with respect to various evaluation disciplines, among which are immediate antecedency (*ia*) and non-pronominal anchors (*na*) (see section 4.6).

For being compatible with the classifiers learned during the training phase, the application version of ROSANA-ML has to employ the identical feature vector signature, i.e. the same Cartesian product of attribute sets to which the individual instances of anaphors and antecedent candidates are mapped. There are further settings, such as the exact way how the antecedent filtering criteria are to be applied, which should be identical during training set generation and application phase (see below).

3. Algorithms and Implementation

The algorithms employed by ROSANA-ML for training data generation and anaphor resolution are immediate descendants of the robust anaphor resolution algorithm underlying the manually designed system ROSANA. Regarding the full details of how robustness is achieved by ROSANA, the reader is referred to the article (Stuckardt 2001). Like its ancestor, ROSANA-ML handles a broad range of entity-specifying expressions, in particular ordinary, possessive, reflexive/

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: apply the robust syntactic disjoint reference filter as specified in (Stuckardt 2001), Figure 4;
- ...
2. *Feature vector generation*: for each remaining anaphor-candidate pair (α_i, γ_j) :
 - a. generate, according to the feature signature σ under consideration, the feature vector $fv(\alpha_i, \gamma_j) := (n_{\alpha_i}, n_{\gamma_j}, f_1, \dots, f_{k_\sigma})$ where n_{α_i} and n_{γ_j} are the number (unique identifiers referred to in the key) of the occurrences α_i and γ_j , and f_1, \dots, f_{k_σ} are (individual and relational) features derived from α_i and γ_j with respect to the signature σ ;
 - b. write $fv(\alpha_i, \gamma_j)$ to the training data file *data.fve*.

Figure 2: ROSANA-ML: training data generation

reciprocal, and relative pronouns, definite NPs, and names. The machine learning experiments described in this article will focus on the key cases of third person non-possessive and possessive *pronominal* anaphora.⁴

In aiming at determining the coreference classes of non-zero linguistic expressions which specify entities⁵, ROSANA-ML covers the coreference task of the Message Understanding Conferences (see (Hirschman 1998; Vilain, Burger, Aberdeen, Connolly & Hirschman 1996; Grishman & Sundheim 1996)).

3.1 *Training data generation*

Figure 2 gives the specification of the training data generation algorithm. The antecedent filtering step 1, in which different kinds of restrictions for eliminating

⁴ As modeled, e.g., by Binding Principle A of the Government and Binding (GB) theory by Chomsky (1981), there are tight syntactic bounds that confine the antecedent options for reflexives and reciprocals. Since these restrictions can be robustly implemented, pronouns of these types can be resolved with very high precision and recall anyway. A similar observation holds with respect to relative pronouns, which, too, can be interpreted with high accuracy by mere surface positional and syntactic means. Hence, these types of anaphoric expressions are not considered during the machine learning experiments; anyway, they are dealt with by appropriate manually designed interpretation strategies as specified in (Stuckardt 2001). Regarding definite NPs and names, machine learning experiments should be based on additional lexico-semantic and ontological information not taken into account in the purely syntactic framework of the ROSANA approach.

⁵ in contrast to expressions that, e.g., specify events

impossible antecedents (in particular, agreement in person/number/gender and syntactic disjoint reference) are applied, is immediately taken over from the original ROSANA algorithm. In step 2, however, no salience ranking of the remaining antecedent candidates is performed. Rather, each remaining anaphor-candidate pair (α_i, γ_j) is mapped to a feature vector $fv(\alpha_i, \gamma_j)$, the attributes $f_1, \dots, f_{k\sigma}$ of which comprise individual and relational features derived from the descriptions of the occurrences α_i and γ_j . The *signature* of the feature vectors, i.e. the inventory of features to be taken into account⁶ has to be chosen carefully in order to fulfill the conditions of robust processing: instead of requiring complete and unambiguous descriptions, they should be computable from potentially partial representations such as *fragmentary* syntactic representations.⁷

As initially motivated, by restricting the consideration to (α_i, γ_j) instances in which γ_j denotes an antecedent candidate that, relatively to α_i , doesn't violate any tight condition, the learning approach focuses on a subset of cases that, from a knowledge engineering point of view, are difficult to decide upon algorithmically (see Figure 3).⁸ In other words, machine learning techniques are applied only for handling the (presumably) difficult cases, i.e. to discern between cospecifying and non-cospecifying candidates that, at current, cannot be distinguished by applying one of the robustly computable restrictions.⁹

⁶ Formally, the feature vectors are instances of an underlying signature, which is defined as the Cartesian product of the sets of attributes taken into account: $fv(\alpha_i, \gamma_j) \in A_1 \times A_2 \times \dots \times A_{k\sigma}$

⁷ The two additional "technical" features n_{α_i} and n_{γ_j} , which correspond to the head token surface number of anaphor and candidate, are required for relating the output *data.fve* to the key data; they are removed during the generation of the file *data.fvc* of classified vectors.

⁸ Of course, as pointed out by Mitkov (1997), the distinction between tight constraints and fuzzy preference criteria is all but uncontroversial. Suffice it to say that what is perceived as a tight constraint is determined through our current state of knowledge: what, today, may be taken as a weak preference criterion, could in future, based on a deeper insight into the problem, be stepwisely refined such that, eventually, a tight criterion emerges.

⁹ A closer analysis reveals that the initially mentioned statistical approaches, too, do at most partially match this clear-cut paradigm. Dagan & Itai (1990) explore a related approach, in which selectional preferences are automatically derived through a statistical analysis of large corpora. Contrary to the methodology followed in the article at hand, they don't make use of coreference information. Importantly, the acquired selectional preference criteria are intended to supplement, rather than substitute, other preference strategies. This has been further explored by Dagan, Justenson, Lappin, Leass & Ribak (1995) and by Lappin & Leass (1994), who showed that, by supplementing a syntactic salience-based anaphor resolver with statistical preferences, an improvement of 2.5 percent can be obtained.

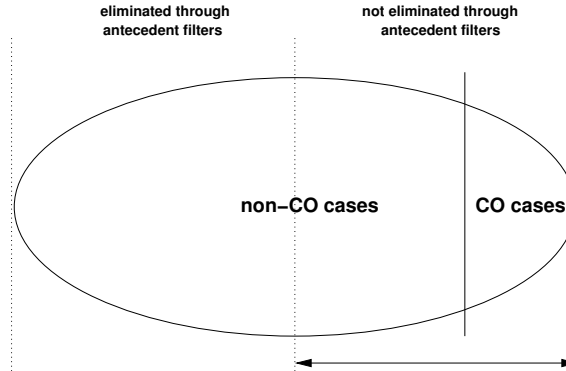


Figure 3: focusing on the relevant cases

3.2 Anaphor resolution

The specification of the ROSANA-ML anaphor resolution algorithm proper is given in Figure 4. Again, step 1 is identical with the antecedent filtering phase of the manually designed ROSANA algorithm. Step 2, however, is modified. For a specific instance (α_i, γ_j) of anaphor and antecedent candidate, after the computation of the feature vector $\text{fv}(\alpha_i, \gamma_j)$, the decision tree lookup takes place; basically, its result $\Psi_\sigma^{\text{type}(\alpha_i)}(\text{fv}(\alpha_i, \gamma_j))$ consists in a prediction $\in \{\text{COSPEC}, \text{NON_COSPEC}\}$.¹⁰

Ge, Hale & Charniak (1998) investigate a statistical approach that employs corpora annotated with syntactic and coreference information to derive antecedent preference criteria. The tree search algorithm developed by Hobbs (1978) is used as the base strategy for anaphor resolution. Thus, in requiring complete syntactic descriptions, this approach cannot be considered to be truly robust, i.e. operational on potentially noisy data. Moreover, while the syntactic disjoint reference conditions can be regarded to be implicitly covered by the tree search algorithm, the filtering criterion of agreement in person/number/gender is left to the responsibility of the statistically derived word-oriented preference criteria. To that extent, their approach differs from the paradigm outlined in Figure 3.

Notably, Paul, Yamamoto & Sumita (1999) investigate the opposite scenario in which decision tree classifiers are employed as candidate *filters* only; since, however, they are studying the case of Japanese restricted domain dialogues, an immediate comparison of their evaluation results with the figures determined below is problematic. Moreover, as initially mentioned, various approaches employ decision tree classifiers as the solitary master strategy, thus implicitly covering antecedent preference and filtering, e.g. (Aone & Bennett 1995; McCarthy & Lehnert 1995; Connolly, Burger & Day 1994).

¹⁰ To put it formally: a *classifier function* $\Psi_\sigma^{\text{type}(\alpha_i)}: A_1 \times A_2 \times \dots \times A_{k_\sigma} \mapsto \{\text{COSPEC}, \text{NON_COSPEC}\}$ is applied that maps instances of the underlying signature σ to cospecification / non-cospecification predictions.

In the subsequent step, these predictions are employed for computing a ranking over the candidate sets of each anaphor. In its base version, candidates which are classified to COSPECify with the anaphor rank higher than candidates that are predicted to NON_COSPECify; surface nearness (i.e. word distance) serves as the secondary criterion. Among the possible refinements are: further ranking the candidates according to the classification error probability yielded by the decision tree lookup, and eliminating candidates which are (fuzzily) classified as NON_COSPECifying (see section 5.5 below, in which results of a series of respective experiments will be given). There is a final step 3 in which the actual antecedent selection takes place. The remaining candidates are considered in the order determined by the ranking procedure; additional means are taken to avoid combinations of antecedent decisions that are mutually incompatible (see (Stuckardt 2001)).

3.3 Implementation

Based on the algorithms described in Figure 2 and Figure 4, the machine-learning-based anaphor resolution system ROSANA-ML has been implemented. Regarding its peripheral modules (definition of basic data structures, preprocessing of the externally provided parsing results, occurrence identification, restriction verification, result scoring), ROSANA-ML is code-identical with its manually designed ancestor. Further components have been added to provide the functionality for training data generation, feature vector classification, decision tree lookup, and modified candidate ranking. The ROSANA-ML System has been implemented in Common Lisp.¹¹

For the task of learning decision tree classifiers from the training data, the C4.5 implementation for Unix of the University of Regina¹² is employed.

¹¹ The FDG parser for English of Järvinen & Tapanainen (1997) has been chosen as the syntactic preprocessor. In giving robustness and processing speed priority over normativeness and syntactic coverage of the underlying grammar, the parser meets the requirements on a preprocessor for robust anaphor resolution on unrestricted texts.

¹² Release 8 for Unix, available (February 1, 2002) at <http://www.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html>

ROLAND STUCKARDT

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: apply the robust syntactic disjoint reference filter as specified in (Stuckardt 2001), Figure 4;
 - ...
2. *Candidate scoring and sorting*:
 - a. for each remaining anaphor-candidate pair (α, γ_j) :
 - i. determine, according to the feature signature σ underlying the learned classifier (decision tree) $\Psi_{\sigma}^{\text{type}(\alpha)}$ to be applied, the internal representation of the feature vector $\text{fv}(\alpha, \gamma_j) := (f_1, \dots, f_{k_{\sigma}})$ where $f_1, \dots, f_{k_{\sigma}}$ are (individual and relational) features derived from the occurrences α_i and γ_j with respect to the applicable signature σ ;
 - ii. *decision tree lookup*: determine the prediction $\Psi_{\sigma}^{\text{type}(\alpha)}(\text{fv}(\alpha, \gamma_j))$ of the learned classifier with respect to the instance $\text{fv}(\alpha, \gamma_j)$.
 - b. for each anaphor α : sort candidates γ_j according the following criteria:
 - i. *primary*: candidates γ_j for which $\Psi_{\sigma}^{\text{type}(\alpha)}(\text{fv}(\alpha, \gamma_j)) = \text{COSPEC}$ are preferred over candidates γ_j for which $\Psi_{\sigma}^{\text{type}(\alpha)}(\text{fv}(\alpha, \gamma_j)) = \text{NON_COSPEC}$;
 - ii. *secondary*: surface nearness.
 - c. sort the anaphors α according to the above criteria applied to 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 choosing $\gamma_j(\alpha)$ as antecedent for α , does neither violate the binding principles nor the i-within-i condition. (see the full specification in (Stuckardt 2001), Figure 4)

Figure 4: ROSANA-ML: anaphor resolution through learned classifiers

4. *Layout of Experiments*

A series of experiments at different levels of consideration will be carried out:

4.1 *Variation of feature vector signatures*

The first and most fundamental question regards the set of attributes, i.e. the signature of the feature vectors from which the classifiers will be learned. As pointed out above, the choice is confined to attributes that are *robustly computable* over the morphological, syntactical, and semantic information available under application conditions. Actual signatures are then defined by selecting subsets of the above attributes.

In Table 1, the set of attributes currently taken into account is shown. $type(o)$ denotes the type of the respective occurrence o , in particular PER3/POS3 (third person non-possessive/possessive pronouns), VNOM (ordinary noun phrases), and NAME (proper names); regarding the anaphor ($o = \alpha$), the choice is restricted to PER3 and POS3 in the current experiments. The feature $synfun(o)$ describes the syntactic function of o . $synlevel(o)$ captures a coarse notion of (non-relational) syntactic prominence¹³, which is measured by counting the number of principal categories¹⁴ occurring on the path between o and the root of the respective parse fragment. Features $number(o)$ and $gender(o)$ capture the respective morphological characteristics of anaphor α and candidate γ . Furthermore, surface context information about the three neighbours to the left and to the right of α and γ is taken into account, comprising the syntactic category ($syncateg(o)$) and, again, the syntactic function ($synfun(o)$) of the respective token(s). Finally, four relational features are considered: $dist(\alpha, \gamma)$ (sentence distance, only distinguishing between three cases: same sentence, previous sentence, two or more sentences away), $dir(\alpha, \gamma)$ (whether γ topologically precedes α or vice versa), $synpar(\alpha, \gamma)$ (identity of syntactic function)¹⁵, and $syndom(\alpha, \gamma)$ (relative syntactic position of the clauses of anaphor α and candidate γ if they occur in the same sentence)¹⁶.

¹³ in contrast to *relational* notions of syntactic prominence, in which the relative position to the other occurrence is taken into account (e.g. *c-command*)

¹⁴ to put it more formally: nodes that, in the sense of the Government and Binding (GB) theory, constitute *binding categories* (see (Chomsky 1981))

¹⁵ thus immediately capturing the role inertia information that has been found to be useful in the classical, manually designed approaches (Lappin & Leass 1994; Stuckardt 2001).

¹⁶ e.g., $[\alpha \rightarrow \gamma]$ describes the case in which the clause of γ is syntactically subordinated to the clause of α

ROLAND STUCKARDT

Feature	Examples of Instances	Description
$\text{type}(\alpha)$	PER3, POS3	type of anaphor α
$\text{synfun}(\alpha)$	subje, trans	syntactic function of α
$\text{synlevel}(\alpha)$	TOP, SUB, SUBSUB	syntactic position of α
$\text{number}(\alpha)$	SG	morphological number of α
$\text{gender}(\alpha)$	MASK	gender of α
$\text{syncateg}(\text{ln}_i(\alpha))$	N, DET	category of left neighbour i , $1 \leq i \leq 3$
$\text{synfun}(\text{ln}_i(\alpha))$	subje, trans	synt. function of left neighbour i , $1 \leq i \leq 3$
$\text{syncateg}(\text{rn}_i(\alpha))$	N, DET	category of right neighbour i , $1 \leq i \leq 3$
$\text{synfun}(\text{rn}_i(\alpha))$	subje, trans	synt. function of right neighbour i , $1 \leq i \leq 3$
$\text{type}(\gamma)$	VNOM, NAME, PER3	type of candidate γ
$\text{synfun}(\gamma)$	subje, trans	syntactic function of γ
$\text{synlevel}(\gamma)$	TOP, SUB, SUBSUB	syntactic position of γ
$\text{number}(\gamma)$	SG	morphological number of γ
$\text{gender}(\gamma)$	MASK	gender of γ
$\text{syncateg}(\text{ln}_i(\gamma))$	N, DET	category of left neighbour i , $1 \leq i \leq 3$
$\text{synfun}(\text{ln}_i(\gamma))$	subje, trans	synt. function of left neighbour i , $1 \leq i \leq 3$
$\text{syncateg}(\text{rn}_i(\gamma))$	N, DET	category of right neighbour i , $1 \leq i \leq 3$
$\text{synfun}(\text{rn}_i(\gamma))$	subje, trans	synt. function of right neighbour i , $1 \leq i \leq 3$
$\text{dist}(\alpha, \gamma)$	INTRA, PREV, PPREV	sentence distance between γ and α
$\text{dir}(\alpha, \gamma)$	ANA, KATA	resumption: anaphoric or cataphoric?
$\text{synpar}(\alpha, \gamma)$	YES, NO	syntactic role identity (parallelism)?
$\text{syndom}(\alpha, \gamma)$	$[\alpha \rightarrow \gamma]$, $[\gamma \rightarrow \alpha]$, none	synt. dominance relations betw. clauses?

Table 1: complete set of features over which the signatures are defined

At the first experimental level, different subsets of attributes will be considered. In particular, it is experimented with signature σ_{full} , which comprises the complete attribute set (38 features), signature σ_{n1} , in which only the syntactic categories and functions of the *immediately* preceding and following neighbours are taken into account (22 features), and signature σ_{n0} , where the *syncateg* and *synfun* attributes of the neighbours are completely ignored (14 features).

4.2 Variation of training data generation settings

At the second experimental level, different settings regarding the extension of the sets of training vectors to be generated are taken into consideration.

Extending the training set by switching off recency limits: in the manually designed algorithm ROSANA, a recency filter is applied that eliminates candidates for pronominal anaphors that are, in terms of sentence distance, too far away. This strategy, which is justified by the observation that, in most cases, antecedent occurrences are available in the immediately preceding sentences,

drastically reduces the amount of generated training data. It is experimentally evaluated what may be gained by switching off this filter during the training data generation phase.

General or specialized classifiers: according to an important result of the formal evaluation of the manually designed system ROSANA, *different* antecedent preference criteria should be employed for third-person non-possessive and possessive pronouns. To reflect this observation, it is experimented with the strategy of generating training data sets for learning two different, i.e. *specialized* classifiers, one of which dedicated for dealing with non-possessives, the other of which designed for handling possessives.

From a learning-theoretical point of view, the so-far sketched variations of the experimental settings are considered to be redundant as long as enough training data are available: the decision-tree learning algorithm should be able to automatically discern between important and unimportant features, and, moreover, should determine for itself whether the classifiers for non-possessive and possessive third-person persons should be kept apart, reflected by the occurrence of the type(α) feature at or near the root of the derived decision tree classifier; a similar argument holds with respect to the extension of the training set by switching off the recency limits, which results in a (supposedly slight) adulterating of the training data since cases are taken into account that may, according to their general characteristics, differ from the general characteristics of the application-relevant cases inside the recency bounds. However, it will turn out that these experimental variations are successful techniques for heuristically coping with the problem of training data sparsity.

4.3 Variation of C4.5 decision tree learning settings

The C4.5 algorithm provides for different settings that determine how the decision trees are learned. One key parameter, the so-called *pruning confidence factor* *CF*, characterizes the amount of *pruning* (given in percent) performed for avoiding overfitting the training data (see Mitchell 1997). The optimal setting of this parameter depends upon the amount and the reliability of the available training data. Using the empirically best signature and training data generation settings as the point of departure, it will be experimented with different values for this factor.

4.4 Intrinsic (decision tree) cross-validation and learning curve analysis

The final output of the learning phase consists of decision tree classifiers. At the level of *intrinsic cross-validation*, it will be evaluated how these classifiers perform with respect to their basic predictions $\in \{\text{COSPEC}, \text{NON_COSPEC}\}$, averaged over ten experiments with varying sets of training and evaluation data. For this 10-fold cross-validation, the set of feature vectors of the training/evaluation corpus is randomly split into 10 parts of equal size. Ten different experiments are run in which decision trees are determined over nine of the ten subsets and evaluated on the remaining subset. The cumulated (average) results are given as *confusion matrices* that describe how the learned classifiers perform with respect to the two classes COSPEC and NON_COSPEC.

Furthermore, for the training set / evaluation set split on which the classifiers yield median results, the *learning curves* of the classifiers will be analyzed, thus, in particular, giving evidence regarding the amount of training data typically needed for obtaining an empirically optimal (classifier-intrinsic) performance.

4.5 Extrinsic (application-level) cross-validation

A similar experiment of (here: 6-fold) cross-validation will be carried out at the application (i.e. anaphor resolution) level. In contrast to the above experiment of classifier-oriented cross-validation, the random split of the training data is performed at the document level, i.e. the overall corpus, which comprises 66 documents, is split into six subsets of eleven documents each. In this case, cross-validation results with respect to the main evaluation disciplines of the anaphor resolution task will be determined.

4.6 Text corpus and disciplines of formal evaluation

The training and evaluation of the ROSANA-ML system will be performed on a corpus of 66 news agency press releases, comprising 24,712 words, 406 third-person non-possessives¹⁷, and 246 third-person possessive pronouns. For the first three experimental stages, the corpus is firmly partitioned into a training subset (31 documents, 11,808 words, 202 non-possessives, 115 possessives) and an evaluation subset (35 documents, 12,904 words, 204 non-possessives, 131 possessives); for the cross-validation stages, further partitions are generated

¹⁷ Relative pronouns are excluded from consideration since they are effectively resolvable with high accuracy by surface-topological means.

randomly (see above). In all experiments, the training data generation and the application of the trained system take place under conditions of potentially noisy data, i.e. without a-priori intellectual correction of orthographic or syntactic errors.

The anaphor resolution performance will be evaluated with respect to two evaluation disciplines: *immediate antecedency* (*ia*) and *non-pronominal anchors* (*na*). In the first-mentioned discipline, an elementary accuracy measure is employed that determines the precision of correct immediate antecedent choices; by further taking into account cases of unresolved anaphors, the respective recall measure is obtained.¹⁸ In the last-mentioned discipline, the performance with respect to the (application-relevant) determination of *non-pronominal* antecedents is evaluated: precision and recall measures are defined in the same way; however, only non-pronominal antecedent candidates are considered. Thus, the anaphor resolution performance is measured according to the tradeoffs (P_{ia}, R_{ia}) and (P_{na}, R_{na}) . For formal definitions and an in-depth discussion of the evaluation measures, the reader is referred to (Stuckardt 2001).

5. Experiments and Empirical Results

5.1 Optimizing the signature and the training data generation settings

In Table 2, the results of the formal, corpus-based evaluation on the *News Agency Press Releases* corpus are summarized. In the upper line, the scores of the manually designed ROSANA system are given. The next three groups of rows display the evaluation results for the signatures σ_{n0} , σ_{n1} , and, respectively, σ_{full} (see section 4.1, first level of experimental variation). Inside these groups, the training data generation settings are varied (second level). In these stages of experimentation, the partition of the corpus into training data and evaluation data remains fixed ($[d_1^{31}, d_{32}^{66}]$).

In the base level experiment of signature variation (rows labeled (1), (2), (3)), non-possessive and possessive pronouns behave nonuniformly: whereas, with growing number of considered neighbours, non-possessives score marginally better, the performance on possessive pronouns slightly deteriorates.

¹⁸ Under the assumption that *all* pronouns are resolved, the precision measure yields results that are immediately comparable with the accuracy figures given in the evaluations of the classical approaches of, e.g., Lappin & Leass (1994) and (Kennedy & Boguraev 1996). By further allowing for unresolved pronouns, (P,R) tradeoffs are obtained that seem to be comparable with the evaluation results that are given by Aone & Bennett (1995).

ROLAND STUCKARDT

	antecedents (P_{ia}, R_{ia})		anchors (P_{na}, R_{na})	
	PER3	POS3	PER3	POS3
ROSANA (manually)	(0.71,0.71)	(0.76,0.76)	(0.68,0.67)	(0.66,0.66)
(1) $\sigma_{n0}, [d_1^{31}, d_{32}^{66}]$	(0.61,0.60)	(0.71,0.71)	(0.54,0.53)	(0.67,0.66)
(1 _{nc}) = (1) \wedge no cataphors	(0.62,0.62)	(0.77,0.77)	(0.57,0.56)	(0.70,0.70)
(1 ^{tc}) = (1) \wedge type(α)-spec. class.	(0.61,0.60)	(0.69,0.69)	(0.56,0.55)	(0.66,0.65)
(1_{nc}^{tc}) = (1^{tc}) \wedge no cataphors	(0.63,0.63)	(0.76,0.76)	(0.60,0.59)	(0.73,0.73)
(1 _{nc} ^{tc+}) = (1 _{nc} ^{tc}) \wedge no recency filt.	(0.63,0.63)	(0.73,0.73)	(0.58,0.58)	(0.63,0.63)
(2) $\sigma_{n1}, [d_1^{31}, d_{32}^{66}]$	(0.62,0.61)	(0.70,0.70)	(0.54,0.54)	(0.65,0.65)
(2+) = (2) \wedge no recency filter	(0.60,0.60)	(0.70,0.70)	(0.52,0.50)	(0.61,0.60)
(2 _{nc}) = (2) \wedge no cataphors	(0.63,0.62)	(0.74,0.74)	(0.57,0.57)	(0.66,0.66)
(2 ^{tc}) = (2) \wedge type(α)-spec. class.	(0.60,0.60)	(0.70,0.70)	(0.56,0.55)	(0.68,0.67)
(2 _{nc} ^{tc}) = (2 ^{tc}) \wedge no cataphors	(0.63,0.63)	(0.73,0.73)	(0.60,0.59)	(0.65,0.65)
(3) $\sigma_{full}, [d_1^{31}, d_{32}^{66}]$	(0.62,0.62)	(0.69,0.69)	(0.55,0.55)	(0.62,0.62)
(3 ^{tc}) = (3) \wedge type(α)-spec. class.	(0.61,0.61)	(0.69,0.69)	(0.57,0.56)	(0.63,0.62)
(3 _{nc} ^{tc}) = (3 ^{tc}) \wedge no cataphors	(0.62,0.62)	(0.75,0.75)	(0.57,0.56)	(0.64,0.64)
(3+) = (3) \wedge no recency filter	(0.60,0.59)	(0.69,0.69)	(0.49,0.49)	(0.57,0.57)
(3 ^{tc+}) = (3+) \wedge type(α)-spec. cl..	(0.62,0.61)	(0.68,0.68)	(0.54,0.53)	(0.64,0.63)
(3 _{nc} ^{tc+}) = (3 ^{tc+}) \wedge no cataphors	(0.62,0.62)	(0.76,0.76)	(0.58,0.57)	(0.68,0.68)

Table 2: evaluation results: signature and settings variation

More importantly, an in-depth qualitative analysis of the typical failure cases concerning the determination of immediate antecedents revealed that a substantial amount of incorrect decisions could have been avoided by dispreferring cataphoric resumptions¹⁹. In the system ROSANA, this negative preference criterion, which is known to promote a good overall antecedent selection performance, is manually encoded as the so-called cataphora malus factor, which is applied during the antecedent scoring phase. ROSANA-ML, however, failed to learn a respective criterion from the training data, which may be attributed to the fact that the cospecification information employed at the learning-relevant level of *individual* antecedent decisions is *inherently symmetrical*.²⁰ This observation gave rise to a

¹⁹ i.e. cases of anaphora with antecedents surface-topologically *following* the anaphor

²⁰ Antecedent selection interdependency comes into play here. If a cataphoric antecedent candidate, which itself embodies an anaphor, is selected prior to being resolved, some antecedent options are ruled out for this occurrence since selecting a candidate that is known to be cospecifying (in particular: the cataphor that already resumes this occurrence) would not yield any new information. It turned out that in a considerable number of such cases, this occurrence then gets wrongly resolved. This is an immediate consequence of the greedy strategy employed during the antecedent selection phase (step 3 of the ROSANA-ML

further variation at the level of feature vector generation settings: *eliminating instances of cataphoric resumption* in the training as well as in the application phase.

The evaluation results illustrate that, under the *no cataphor* setting, with only one minor exception, results improve considerably. In particular, this holds for possessive pronouns: in experiment (1_{nc}), e.g., the gain in the *immediate antecedency* discipline amounts to 6 points of percentage for P_{ia} and R_{ia} each; in the *nonpronominal anchor* discipline, the improvement is reflected too, amounting to 3% for P_{na} , and 4% for R_{na} .

training set generation settings	training set sizes		
	general	PER3	POS3
standard	7,696	4,804	2,892
no cataphors	7,116	4,446	2,670
no recency filter	17,416	11,115	6,301
no cataphors, no recency filter	16,836	10,757	6,079

Table 3: sizes of the training sets

Extending the training set by switching off the recency limits seems to induce, at first sight, a deterioration: compare, e.g., experiments (1_{nc}^{tc}) and (1_{nc}^{tc+}), or (2) and (2+). However, the comparison of the case series [(3), (3^{tc}), (3_{nc}^{tc})] vs. [(3+), (3^{tc+}), (3_{nc}^{tc+})] shows that this observation doesn't generalize. Rather, it seems to depend on the further settings: in the last-mentioned case, in which the *no cataphor* as well as the *type-specific classifier* settings are activated, there is a slight gain with respect to immediate antecedency of possessive pronouns, and a slight to considerable gain concerning the *nonpronominal anchors* scores for non-possessives and possessives. This may be explained by referring to the respective training set sizes, which are displayed in Table 3. In the base ("standard") case (3), one general classifier is constructed over 7,696 vectors. In the *type-specific classifier* setting, two specialized classifiers have to be learned, the one for non-possessives over 4,804 samples, the one for possessives over 2,892 samples. Under the *no cataphor* setting, the respective training set sizes are further reduced to 4,446 and 2,670, respectively. The observation may thus be explained by referring to the argument of section 4.2: if the amount of available data is sufficiently large, the adulterating effect of artificially enlarging the training set prevails; if, however, training data are sparse, the overall effect may be positive.

algorithm, see Figure 4), which, in order to avoid exponential time complexity, doesn't optimize the *combined* plausibility of the antecedent decisions.

		antecedents (P_{ia}, R_{ia})		anchors (P_{na}, R_{na})	
		PER3	POS3	PER3	POS3
(1_{nc}^{tc})	(CF = 25%)	(0.63,0.63)	(0.76,0.76)	(0.60,0.59)	(0.73,0.73)
$(1_{nc}^{tc}, 15) = (1_{nc}^{tc}) \wedge CF = 15\%$		(0.63,0.62)	(0.76,0.76)	(0.61,0.60)	(0.69,0.69)
$(1_{nc}^{tc}, 37) = (1_{nc}^{tc}) \wedge CF = 37\%$		(0.65,0.64)	(0.72,0.72)	(0.61,0.61)	(0.64,0.64)
$(1_{nc}^{tc}, 50) = (1_{nc}^{tc}) \wedge CF = 50\%$		(0.63,0.62)	(0.72,0.72)	(0.56,0.56)	(0.62,0.62)
$(1_{nc}^{tc}, 62) = (1_{nc}^{tc}) \wedge CF = 62\%$		(0.62,0.61)	(0.72,0.72)	(0.55,0.55)	(0.61,0.61)
$(1_{nc}^{tc}, 75) = (1_{nc}^{tc}) \wedge CF = 75\%$		(0.62,0.62)	(0.72,0.72)	(0.56,0.56)	(0.61,0.61)
$(1_{nc}^{tc}, h) = (1_{nc}^{tc}) \wedge CF = (37 25)\%$		(0.65,0.64)	(0.76,0.76)	(0.62,0.61)	(0.73,0.73)

Table 4: evaluation results: pruning confidence factor variation

The *type-specific classifiers* setting yields nonuniform effects. In some cases, there are gains as well as losses ((1) vs. (1^{tc}) , (2) vs. (2^{tc})). As identified above, however, specialized classifiers seem to pay off in combination with the *extended training set* mode. A particular behaviour is exhibited by the (1_{nc}^{tc}) experiment, which, in terms of overall (averaged) performance, can be considered to comprise the empirically optimal settings: whereas, concerning signature σ_{n0} , the *type-specific classifier* setting alone doesn't yield an overall positive contribution ((1^{tc}) vs. (1)), together with the *no cataphor* setting, the positive effects prevail. In this specific case, the advantage of employing specialized classifiers may outweigh the disadvantage of the small number of training cases since the number of attributes of signature σ_{n0} is considerably lower than in the case of σ_{full} (14 vs. 38).

Through further experiments, the results of which are not displayed in Table 2, the positive contributions of various subclasses of features have been validated. E.g. the evaluation of a signature σ_{full}^{synpro} , which consists of the features of σ_{full} minus the *synlevel* and *syndom* attributes (35 features), confirmed the positive contribution of the non-relational and relational attributes of syntactic prominence.

5.2 Optimizing the C4.5 decision tree learning settings

The settings of the experiment (1_{nc}^{tc}) have been taken as the starting point of further variations at the level of the C4.5 decision tree learning proper (see section 4.3), viz. different settings of the *pruning confidence factor* CF . The base value of CF in all above-discussed experiments was 25 percent. Hence, it has been experimented with further CF values of 15, 37, 50, 62, and 75%. For possessive pronouns, according to the respective results, which are given in Table 4, the original setting of $CF=25\%$ yields the best scores; classifiers for non-possessives, however, should be determined with a slightly higher CF of 37%. Again, this may

be explained by the different sizes of the training sets: for non-possessives, more training cases are available, resulting in a decision tree that better generalizes, thus necessitating a lower amount of pruning, i.e. allowing for a higher pruning confidence factor.

The row $(1_{nc}^{tc}, h)$ displays the evaluation results of a *hybrid setting* in which specialized classifiers for non-possessives and possessives are computed with the respective empirically optimal choices of CF values.

cases	CO	\neg CO	cases	CO	\neg CO
CO	62.7%	37.3%	CO	59.4%	40.6%
n = 1,518	952	566	n = 1,066	633	433
\neg CO	3.5%	96.5%	\neg CO	4.0%	96.0%
n = 8,187	284	7,903	n = 4,777	190	4,587

Table 5: 10-fold cross-validation, confusion matrices (signature σ_{n0} , experiment $(1_{nc}^{tc}, h)$): PER3 and POS3 classifiers

5.3 Intrinsic cross-validation and learning curves

In Table 5, the results of a *10-fold intrinsic cross-validation* according to the method outlined in section 4.4 are given. The two confusion matrices display the overall (cumulated) scores for the PER3 and POS3 classifiers that have been derived by employing the settings of experiment $(1_{nc}^{tc}, h)$.

Regarding the PER3 classifier, the overall number of training/evaluation vectors amounts to 9,705, of which 1,518 are COSPEC instances and 8,187 are NON_COSPEC instances. According to the upper row of the PER3 table, 62.7% of the COSPEC cases are correctly classified, and 37.3% are erroneously classified as belonging to the NON_COSPEC class. The scores for the NON_COSPEC vectors, which are shown in the lower row, are considerably higher: 96.5% correct, 3.5% incorrect. The results obtained for the POS3 classifier are similar. Here, the overall number of training/evaluation vectors is lower (5,843). Of the 1,066 COSPEC cases, 59.4% are correctly classified; instances of the 4,777 vectors belonging to the NON_COSPEC class are identified with an accuracy of 96%.

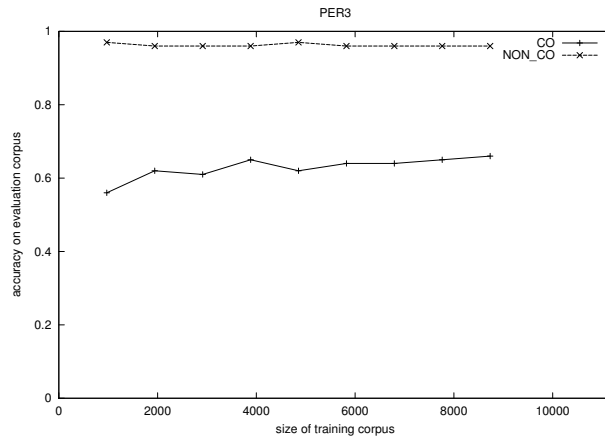


Figure 5: learning curve, signature σ_{n_0} , experiment (I_{nc}^{tc}, h) : PER3 classifier

At first sight, the comparatively low accuracy obtained for the COSPEC cases seems to impose a problem. The situation is not as worse as it looks like since, with respect to step 2b of the ROSANA-ML algorithm (see Figure 4), it is of primary importance not to misclassify the NON_COSPEC cases; wrongly classifying a COSPEC instance is unproblematic as long as there are further cospecifying antecedent candidates that are correctly recognized. A closer analysis shows that, for most anaphors, indeed, *several* cospecifying candidates are available. However, one also has to take in account that the relative amount of NON_COSPEC instances is quite high.²¹ Hence, although the NON_COSPEC instance classification error rate lies clearly below 5%, the probability that, for a certain anaphor to be resolved, one of the (typically numerous) non-cospecifying antecedent candidates gets wrongly classified is still not neglectable.

Out of the random 10-fold partition of the training data, the respective subsets for which median scores were obtained during cross-validation have been employed as the target data for a *learning curve analysis* of the two classifiers. In Figure 5, the learning curve of the PER3 classifier is shown (1,378 CO + 7,356 NON_CO = 8,734 training cases, 140 CO + 831 NON_CO = 971 evaluation cases). It turns out that, for obtaining a classifier that achieves a high performance regarding NON_COSPEC instances, only a small amount of training cases is needed, whereas, regarding the COSPEC cases, a corpus of at least 4,000 sample vectors is necessary for obtaining a performance near the level that was

²¹ see Table 5: about 84% of all instances for PER3, and about 82% of all instances for POS3

A MACHINE LEARNING APPROACH TO PREFERENCE STRATEGIES

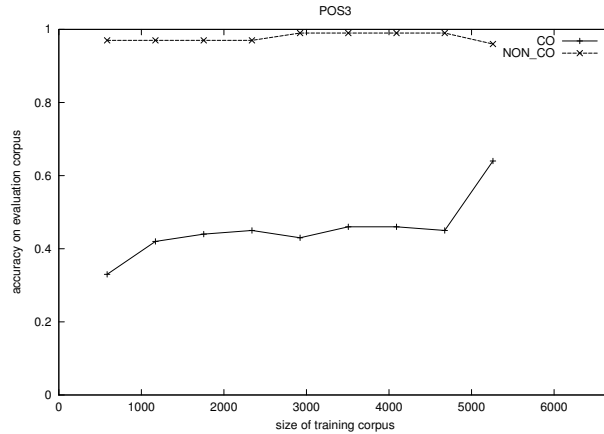


Figure 6: learning curve, signature σ_{n_0} , experiment (l_{nc}^{tc}, h) : POS3 classifier

empirically observed during cross-validation.²² Regarding the POS3 classifier (957 CO + 4,302 NON_CO = 5,259 training cases, 109 CO + 475 NON_CO = 584 evaluation cases), the situation is similar (see Figure 6). Hence, in particular, the COSPEC case recognition curve illustrates that training data sparsity is an issue: at least a training corpus of size 5,000 (1,000 COSPEC instances) seems to be needed for approaching the 60% level that was empirically observed to be achievable.

5.4 Extrinsic (application-level) cross-validation

The results of a 6-fold cross-validation at the application (anaphor resolution) level are displayed in Table 6. According to the method outlined in section 4.5, the data have been randomly split into six subsets d_{si} , $1 \leq i \leq 6$, of eleven documents each. Hence, there are six base experiments with differing training set / evaluation set assignments, viz. $[d_1^{66} \setminus d_{si}, d_{si}]$, $1 \leq i \leq 6$.

²² Partly, this may be regarded a consequence of the fact that the training corpus contains about five times more NON_COSPEC instances than COSPEC cases (7,356 vs. 1,378). However, even upon a respective rescaling of the x-axis, still the NON_COSPEC recognition accuracy curve reaches its empirical optimum faster.

experiment	antecedents (P_{ia}, R_{ia})		anchors (P_{na}, R_{na})	
	PER3	POS3	PER3	POS3
$(1_{nc}^{tc}, h)$ $[d_1^{31}, d_{32}^{66}]$, cf. Table 4	(0.65,0.64)	(0.76,0.76)	(0.62,0.61)	(0.73,0.73)
(ds1) $[d_1^{66} \setminus d_{s1}, d_{s1}]$	(0.71,0.70)	(0.90,0.90)	(0.67,0.65)	(0.79,0.79)
(ds2) $[d_1^{66} \setminus d_{s2}, d_{s2}]$	(0.59,0.59)	(0.70,0.70)	(0.51,0.51)	(0.59,0.59)
(ds3) $[d_1^{66} \setminus d_{s3}, d_{s3}]$	(0.72,0.72)	(0.72,0.72)	(0.73,0.72)	(0.70,0.70)
(ds4) $[d_1^{66} \setminus d_{s4}, d_{s4}]$	(0.82,0.82)	(0.80,0.80)	(0.82,0.82)	(0.74,0.74)
(ds5) $[d_1^{66} \setminus d_{s5}, d_{s5}]$	(0.59,0.59)	(0.76,0.76)	(0.53,0.53)	(0.70,0.70)
(ds6) $[d_1^{66} \setminus d_{s6}, d_{s6}]$	(0.52,0.52)	(0.69,0.69)	(0.45,0.45)	(0.56,0.56)
(ds1-6) cumulated / averaged	(0.66,0.66)	(0.75,0.75)	(0.62,0.62)	(0.68,0.68)

Table 6: 6-fold cross-validation of anaphor resolution results

Regarding the results of the six base experiments, the variance is considerable. Similar observations have been made during the evaluation of the manually designed ROSANA system. Thus, rather than indicating a specific problem of the machine learning approach, the variance seems to be determined by the individual empirical difficulty of the document sets with respect to the anaphor resolution task. With the exception of the nonpronominal anchors result for possessives, which is lower (-5%), the cumulated score (ds1-6) lies close to the figures determined in the $(1_{nc}^{tc}, h)$ experiment. One might expect that, since the training sets are considerably larger than in the original $[d_1^{31}, d_{32}^{66}]$ experiment (on average, (8,734;5,259) vs. (4,446;2,670)), results should be better, particularly for possessive pronouns. One should, however, keep in mind that the learning characteristics of the classifiers are only indirectly mirrored in the anaphor resolution performance (see the discussion in section 5.3); in particular, this holds for the secondary discipline of nonpronominal anchor determination. Hence, though it should certainly be instructive to re-run the experiments on larger data sets, the results of the extrinsic (application-level) cross-validation can be interpreted as confirming the order of magnitude of the figures obtained in the original experiment $(1_{nc}^{tc}, h)$.

5.5 Trading off recall for precision

Based on the empirically optimal configuration $(1_{nc}^{tc}, h)$, a series of further experiments has been carried out to address the question whether, by looking at the additional quantitative information given at the leaves of the C4.5 decision trees, it is possible to gradually bias ROSANA-ML towards *high precision anaphor resolution*. Each decision tree leaf provides the total number μ of *training cases* that match the respective decision path, and the number $\varepsilon \leq \mu$ of these cases

A MACHINE LEARNING APPROACH TO PREFERENCE STRATEGIES

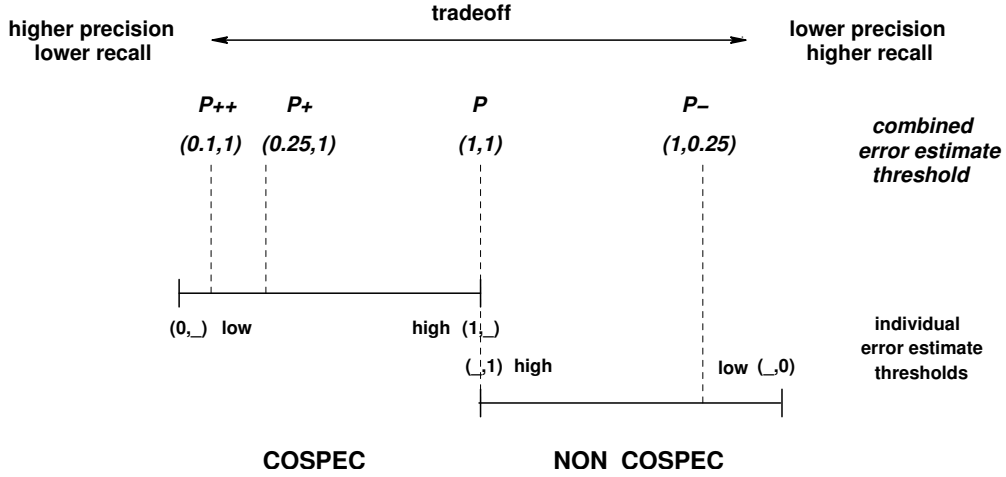


Figure 7: combined error estimate threshold for trading off recall for precision

that are, through the category prediction of the leaf, *misclassified*. By computing the quotient ε/μ , it should thus be possible to derive an estimate of the classification error probability of the specific leaf.

This information can now be used to gradually bias ROSANA-ML towards high precision anaphor resolution. The base version of the algorithm specified in Figure 4 prefers candidates predicted to COSPECify over candidates predicted to NON_COSPECify, and employs surface-topological distance as the secondary criterion. By looking at the quotient ε/μ , this preference criterion may be refined as follows: prefer COSPEC candidates over NON_COSPEC candidates; at the secondary level, prefer COSPEC candidates with smaller classification error estimate ε/μ over COSPEC candidates with higher ε/μ , and prefer NON_COSPEC candidates with higher classification error estimate ε/μ over NON_COSPEC candidates with lower ε/μ . Finally, by setting a *threshold* $\theta := (\theta_{co}, \theta_{nonco})$ as illustrated in Figure 7, 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* (\leq) θ_{nonco} , a bias may be imposed that gradually trades off recall for precision.

experiment	antecedents (P_{ia}, R_{ia})		anchors (P_{na}, R_{na})	
	PER3	POS3	PER3	POS3
$(1_{nc}^{tc}, h) [d_1^{31}, d_{32}^{66}]$, cf. Table 4	(0.65,0.64)	(0.76,0.76)	(0.62,0.61)	(0.73,0.73)
$(1_{nc}^{tc}, h, p) = (1_{nc}^{tc}, h) \wedge \theta = (1.0, 1.0)$	(0.79,0.51)	(0.86,0.60)	(0.75,0.45)	(0.83,0.54)
$(1_{nc}^{tc}, h, p-) = (1_{nc}^{tc}, h) \wedge \theta = (1.0, 0.25)$	(0.74,0.56)	(0.78,0.63)	(0.71,0.52)	(0.76,0.59)
$(1_{nc}^{tc}, h, p+) = (1_{nc}^{tc}, h) \wedge \theta = (0.25, 1.0)$	(0.81,0.45)	(0.89,0.50)	(0.74,0.36)	(0.67,0.30)
$(1_{nc}^{tc}, h, p++) = (1_{nc}^{tc}, h) \wedge \theta = (0.1, 1.0)$	(0.83,0.31)	(1.00,0.17)	(0.80,0.08)	(1.00,0.12)

Table 7: evaluation results: trading off recall for precision

In Table 7, the results of four respective experiments with different threshold settings are displayed. The basic precision bias setting of experiment $(1_{nc}^{tc}, h, p)$ allows ε/μ values of ≤ 1 for COSPEC-predicted instances, and > 1 for NON_COSPEC-predicted instances; in other words, all and only the candidates that are predicted not to cospecify are eliminated. 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 $\theta < 1$, e.g. $\theta = 0.25$ (experiment $(1_{nc}^{tc}, h, p-)$). Similarly, the bias can be strengthened by imposing lower error ratio thresholds for the candidates predicted to COSPECify (experiments $(1_{nc}^{tc}, h, p+)$ and $(1_{nc}^{tc}, h, p++)$). Regarding the primary evaluation discipline of immediate antecedency, the scores for the different settings indicate that, as expected, by employing the quantitative information given at the decision tree leaves in the above-described way, one obtains a suitable (albeit heuristic) means for gradually biasing ROSANA-ML towards high precision. Regarding the nonpronominal anchor discipline, which is more indirectly related to the classifier predictions, this conjecture can be regarded to be confirmed.

6. Comparison

6.1 ROSANA-ML vs. ROSANA

The comparison of the evaluation results for experiment $(1_{nc}^{tc}, h)$ (see Table 4) with the scores of the manually designed ROSANA system on $[d_1^{31}, d_{32}^{66}]$ (see Table 2) leads to a nonuniform assessment. Whereas ROSANA-ML performed better with respect to nonpronominal anchors for possessive pronouns, the results for non-possessives deteriorated. At first sight, this is surprising since, as observed in section 5.3, regarding the classifier for possessives, the training set size of 2,670 is too small to arrive at the possible accuracy level of around 60% with respect to

the recognition of COSPEC cases. However, one has to take into account that, according to the results that have been determined for the manually designed ROSANA system, possessives are generally easier to resolve than non-possessives.

In view of the efforts that went into the refinement of the preference factors employed in manually designed systems, the results can be regarded to be encouraging. With a comparatively low amount of training data, a performance regarding possessives has been achieved that at least reaches, if not outperforms the results of the hand-tuned ROSANA approach.²³ The inferior results on non-possessives can be interpreted as an indicator that the inventory of feature sets over which the signatures are defined should be enlarged. According to the learning curve analysis in section 5.3, at least in the extrinsic (application-level) cross-validation experiments, the training set size should have been sufficiently large ($> 8,000$) to arrive at the possible accuracy level of around 60% with respect to the recognition of COSPEC cases. This gives evidence that, for arriving at an anaphor interpretation performance on non-possessives similar to the performance of manually designed systems, a COSPEC accuracy of 60% does not suffice, and, moreover, that yet not a sufficient inventory of features (Table 1) is available.

6.2 ROSANA-ML vs. Aone and Bennett (1995)

In their machine learning approach to anaphor resolution of Japanese texts, Aone & Bennett (1995) determine (P,R) figures regarding four types of anaphoric expressions: names, definite NPs, quasi-zero pronouns, and zero pronouns. The investigation is restricted to anaphoric expressions that specify organizations. Hence, their findings do not immediately compare with the evaluation results given above, which have a more general scope. A first, coarse impression, however, may be obtained by comparing the results regarding possessive and non-possessive pronouns with the cases of Japanese quasi-zero and zero pronouns, for which Aone and Bennett give immediate antecedency figures of $(0.85,0.64)$ and $(0.76,0.38)$. Under the assumption that similar definitions of the precision and

²³ According to the results of the 6-fold extrinsic (application-level) cross-validation given in Table 6, the performance regarding possessives is expected to be, on average, lower than observed for the $(1_{nc}^{tc},h)$ experiment on $[d_1^{31},d_{32}^{66}]$. At least the level of the ROSANA scores on $[d_1^{31},d_{32}^{66}]$ has been reached; however, a more instructive comparison should be based on data obtained from a comparable in-depth cross-validation of ROSANA.

recall measures are employed,²⁴ these results can be compared to the scores of the high precision anaphor resolution experiments that are summarized in Table 7. Whereas the quasi-zero pronoun figures (0.85,0.64) seem to indicate (at least when compared to the immediate antecedency scores for non-possessives) that the Aone & Bennett (1995) approach outperforms ROSANA-ML, evidence is to the contrary if one takes the zero pronoun figures (0.76,0.38) as the base of comparison. As pointed out by Aone and Bennett, quasi-zero pronouns are more easy to resolve since, by definition, they always cospecify with a local subject, and, hence, may be interpreted by purely syntactical means. Consequently, the zero pronoun scores can be regarded as the more suitable reference for comparison, thus urging upon the conclusion that the methodology of ROSANA-ML, according to which machine learning is applied to derive anaphor resolution *preference* strategies, is superior to the unfocused learning approach employed by Aone & Bennett (1995), in which preferences as well as restrictions are learned.

6.3 ROSANA-ML vs. CogNIAC

Baldwin (1997) describes the CogNIAC approach that achieves high precision coreference resolution by restricting antecedent decisions to cases in which no world knowledge or sophisticated linguistic processing seems to be needed for successful resolution. The recognition of such cases is performed by a set of six manually designed rules. The resolution of only those pronouns is tried the interpretation context of which matches one of these rules; all other pronouns remain unresolved. While it remains unclear whether the employed formal (*P,R*) measures neatly match up with the evaluation criteria used above, the evaluation figures of (0.92,0.64), which were obtained on a corpus with 298 cases of English third person pronouns, seem to give evidence that the manual design of a high precision rule set outperforms the machine-learning-based approach which has been obtained above as a side-product by referring to quantitative information available at the decision tree leaves. However, it has to be taken into account that Baldwin (1997) manually corrected the preprocessing results in order to allow for a fair comparison of his approach with the non-robust algorithm of Hobbs (1978), which employs complete and unambiguous parses. In fact, results of recent experiments indicate that, on potentially noisy data and without intellectual intervention, the ROSANA-ML approach to high precision anaphor resolution at

²⁴ Aone and Bennett (1995) do not give formal definitions of the employed measures; however, there is clear evidence that the measures are equivalent, or nearly equivalent, to the measures used in the article at hand.

least performs on a par with a *robust* reimplementation of Baldwin's algorithm (see Stuckardt (2003)).

Moreover, it remains to be investigated whether even better (P,R) tradeoffs are obtained if decision trees that have been learned over larger amounts of training data are employed. Classifiers of higher quality are expected to yield better estimates of the classification error probability as defined in section 5.5.

7. Conclusion and Further Research

Overall, the evaluation results of ROSANA-ML are promising. According to the above experiments, it can be concluded that, by employing a machine learning approach to preference strategies for anaphor resolution, results that at least compare with those of the best manually tuned systems can be reached. With respect to the current best settings and regarding possessive third person pronouns, the resolution quality is slightly higher than for the ancestor system ROSANA, whereas, regarding non-possessives, the quality slightly lags behind. The cross-validation scores range from 75% (possessives, immediate antecedency) to 62% (non-possessives, nonpronominal anchors). Moreover, the investigation has given evidence that, by biasing ROSANA-ML towards precision, better (P,R) tradeoffs can be obtained than those achieved by the approach of Aone & Bennett (1995). While this can be interpreted as an indicator that the approach employed by ROSANA-ML, which focuses on machine learning *preferences*, may lead to a better overall performance, it has to be kept in mind that the cases of English third-person pronouns and Japanese zero-pronouns do not immediately compare.

Future efforts should focus on the goal of enhancing the interpretation quality regarding non-possessives. According to the results of the above evaluation, most certainly this will require that the set of features over which the classifiers are learned is appropriately supplemented. Finding suitable candidate features, however, can be considered to be a hard and time-consuming intellectual task, thus illustrating that machine learning approaches do not generally free the knowledge engineer from intellectual fine-tuning. In this specific case, the task may be immediately compared with the intellectual determination of suitable robustly computable salience factors; however, the application of decision-tree learning saves part of the time necessary for optimizing the playing together of the overall set of factors in the classical approaches to anaphor resolution.

Larger sets of features over which classifiers are learned will enhance the need for bigger training corpora. This key issue should be considered for further

reasons. First, as outlined in section 4, parts of the first two levels of experimental consideration will become obsolete: if enough training data are available, C4.5 will be able to discover for itself which features are key and which not, thus freeing the knowledge engineer from experimenting with subset signatures, or from artificially enlarging the set of training cases. Moreover, of paramount importance is the availability of sufficiently large corpora of *different text genres*, which is the enabling condition for empirically addressing the issue of genre-specific preference strategy assignment, a goal that has been put forward as a consequence of the evaluation results of the manually designed ROSANA system.

Based on these further experiments on larger and heterogeneous corpora, the learned classifiers should undergo a thorough *qualitative* analysis. Which features do typically occur at or near the root of the learned decision trees? Which features are typically eliminated during pruning? Are there certain characteristics that are specific to the different training corpus genres? Regarding the qualitative exploration of classifiers, it should be worthwhile to look at C4.5 classifiers in the *lists of rules* format, which are generated by the classifier learning tool *C4.5rules* of the employed C4.5 implementation (see section 3.3). From the point of view of knowledge engineering, besides having available enhanced pruning options, an important advantage of employing rules instead of trees lies in their better intellectual accessibility. The qualitative analysis of classifiers might ultimately shed new light on the empirical foundation of classical strategies for determining salience, including theories of attentional focusing such as centering.

References

- Aone, Chinatsu & Scott William Bennett. 1995. "Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies". *Proceedings of the 33rd Annual Meeting of the ACL, Santa Cruz, New Mexico*, 122-129.
- . 1996. "Applying Machine Learning to Anaphora Resolution". *Connectionist, statistical and symbolic approaches to learning for Natural Language Processing*, ed. by S. Wermter, E. Riloff & G. Scheler, 302-314. Berlin: Springer.
- Baldwin, Breck. 1997. "CogNIAC: High Precision Coreference with Limited Knowledge and Linguistic Resources". *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphor Resolution for Unrestricted Texts, Madrid, July 1997*, ed. by Ruslan Mitkov & Branimir Boguraev, 38-45.

A MACHINE LEARNING APPROACH TO PREFERENCE STRATEGIES

- Chomsky, Noam. *Lectures on Government and Binding*. 1981. Dordrecht: Foris.
- Connolly, Dennis, John D. Burger & David S. Day. 1994. "A Machine Learning Approach to Anaphoric Reference". *Proceedings of the International Conference on New Methods in Language Processing (NEMLAP)*.
- Dagan, Ido & Alon Itai. 1990. "Automatic Processing of Large Corpora for the Resolution of Anaphora References" *Proceedings of the 13th International Conference on Computational Linguistics (COLING), Helsinki*, vol. III, 330-332.
- , John Justenson, Shalom Lappin, Herbert Leass & Amnon Ribak. 1995. "Syntax and Lexical Statistics in Anaphora Resolution". *Applied Artificial Intelligence* 9:6, 633-644.
- Ge, Niyu, John Hale & Eugene Charniak. 1998. "A Statistical Approach to Anaphora Resolution". *Proceedings of the Sixth Workshop on Very Large Corpora, Montreal*, 161-170.
- Grishman, Ralph & Beth Sundheim, 1996. "Design of the MUC-6 Evaluation". *Proceedings of the Sixth Message Understanding Conference (MUC-6), 1-11*. Morgan Kaufmann..
- Hirschman, Lynette. 1998. "MUC-7 Coreference Task Definition, Version 3.0.". *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Published online, available (December 9, 1999) at http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html
- Hobbs, Jerry R. 1978. "Resolving Pronoun References". *Lingua* 44. 311-338.
- Järvinen, Timo & Pasi Tapanainen. 1997. *A Dependency Parser for English*. Technical Report TR-1. University of Helsinki: Department of General Linguistics.
- Kennedy, Christopher & Branimir Boguraev. 1996: "Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser". *Proceedings of the 16th International Conference on Computational Linguistics (COLING), Copenhagen*, vol. I, 113-118.
- Lappin, Shalom & Herbert J. Leass. 1994. "An Algorithm for Pronominal Anaphora Resolution". *Computational Linguistics* 20:4. 535-561.
- McCarthy, Joseph F. & Wendy G. Lehnert. 1995. "Using Decision Trees for Coreference Resolution". *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95), Montreal*.

- Mitchell, Tom M. 1997. *Machine Learning*. New York: McGraw-Hill.
- Mitkov, Ruslan. 1997. "Factors in Anaphora Resolution: They are not the Only Things that Matter. A Case Study Based on Two Different Approaches". *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphor Resolution for Unrestricted Texts, Madrid, July 1997*, ed. by Ruslan Mitkov & Branimir Boguraev, 14-21.
- . 2001. "Outstanding Issues in Anaphora Resolution". *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Mexico City*, ed. by Alexander Gelbukh, 110-125.
- Paul, Michael, Kazuhide Yamamoto & Elichiro Sumita. 1999. "Corpus-Based Anaphora Resolution Towards Antecedent Preference". *Proceedings of the ACL'99 Workshop on Coreference and its Applications, Maryland*, 47-52.
- Quinlan, John Ross. 1993. *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.
- Soon, Wee Meng, Hwee Tou Ng & Daniel Chung Yong Lim. 2001. "A Machine Learning Approach to Coreference Resolution of Noun Phrases". *Computational Linguistics* 27:4. 521-544.
- Stuckardt, Roland. 1997. "Resolving Anaphoric References on Deficient Syntactic Descriptions". *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphor Resolution for Unrestricted Texts, Madrid, July 1997*, ed. by Ruslan Mitkov & Branimir Boguraev, 30-37.
- . 2001. "Design and Enhanced Evaluation of a Robust Anaphor Resolution Algorithm". *Computational Linguistics* 27:4. 479-506.
- . 2003. "Coreference-Based Summarization and Question Answering: a Case for High Precision Anaphor Resolution". *Proceedings of the 2003 International Symposium on Reference Resolution and Its Application to Question Answering and Summarization (ARQAS), Università Ca' Foscari, Venice*, 33-41.
- Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly & Lynette Hirschman. 1996. "A Model-Theoretic Coreference Scoring Scheme". *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, 45-52. Morgan Kaufmann.