In: Proc. 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2002), University of Lisbon, Sept. 2002, 211-216.

Machine-Learning-Based vs. Manually Designed Approaches to Anaphor Resolution: the Best of Two Worlds

Roland Stuckardt

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

Abstract

In the last 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 techniques for deriving strategies automatically. Since the knowledge-engineering effort for designing and optimizing the strategies is reduced, the latter approaches are considered particularly attractive. Since, however, the hand-coding of robust antecedent filtering strategies such as syntactic disjoint reference and agreement in person, number, and gender constitutes a once-for-all effort, the question arises whether at all they should be derived automatically.

In this paper, it is investigated what might be gained by combining the best of two worlds: designing the universally valid antecedent filtering strategies manually, in a once-for-all fashion, and deriving the (potentially genre-specific) antecedent selection strategies automatically by applying machine-learning techniques. An anaphor resolution system ROSANA-ML, which follows this paradigm, is designed and implemented. Through a series of formal evaluations, it is shown that, while exhibiting additional advantages, ROSANA-ML reaches a performance level that compares with the performance of 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 years, much effort went into the design of robust algorithms which work under knowledge-poor application conditions. Many approaches take as a starting point the landmark work of Lappin and 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 and 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 techniques for deriving anaphor resolution strategies automatically (Aone and Bennett, 1996; Aone and Bennett, 1995; Connolly et al., 1994). These approaches are considered particularly attractive because the effort for designing and implementing the strategies is reduced. However, deriving anaphor resolution strategies automatically relies upon the availability of sufficiently large text corpora that are tagged, in particular, with referential information.[1]

Aone and Bennett (1995) 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 descent from the work of Lappin and Leass (1994) are, with respect to the robust operationalization of the antecedent filtering strategies[2], nearly optimal (Stuckardt, 2001; Kennedy and 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 machine-learning techniques.

In this paper, it is investigated what might be gained by employing machine-learned *preference* (antecedent selection) strategies as part of a robust anaphor resolution approach according to the Lappin and 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 paper have shown that, for achieving optimal interpretation results, the antecedent selection strategies[3] should be designed in a genre-specific way, since text genres seem to differ w.r.t. the characteristic properties of their typical coherence structures. Hence, there is no once-for-all optimal design of preference heuristics. Consequently, these antecedent selection strategies are ideal candidates for applying machine-learning techniques.

Thus, it is explored what might be gained by combining the best of two worlds: designing the universally valid antecedent filtering strategies manually, in a once-for-all fashion, and deriving the corpus-specific antecedent selection strategies automatically by applying machine-learning techniques. An anaphor resolution system ROSANA-ML, which follows this paradigm, will be designed and implemented. Through a series of formal evaluations, it will be shown that, w.r.t. two important evaluation measures, ROSANA-ML reaches a level of performance that com-

[1] While some referentially annotated corpora have been developed during the last years (particularly for the DARPA Message Understanding Conferences (MUCs)), the total amount of available tagged texts is still quite restricted.

[2] syntactic disjoint reference and number/gender agreement

[3] implemented through a set of *salience factors*

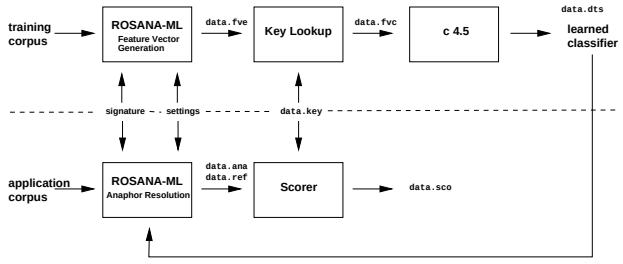

Figure 1: ROSANA-ML: training vs. application case

pares with the interpretation quality of its manually designed ancestor ROSANA. More specifically, the evaluation shows 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 will enhance the need for still larger corpora of referentially-annotated training texts, thus confirming similar findings that have been made elsewhere (e.g. (Mitkov, 2001)). Moreover, the results of a series of further experiments will indicate that, regarding third-person pronominal anaphora in English, by biasing ROSANA-ML towards precision, better precision/recall tradeoffs may be obtained than those determined by Aone and Bennett (1995) for the case of Japanese zero pronouns.

2. Methodology

In figure 1, the machine learning approach to anaphor resolution followed by ROSANA-ML is outlined. It is distinguished between the **training case**, which is shown in the upper part of the figure, and the **application (i.e. anaphor resolution) case** sketched in the lower part of the figure. During the training phase, C4.5 decision tree classifiers are constructed which yield predictions whether, given a pair of anaphor and antecendent candidate, these two occurrences are cospecifying or non-cospecifying; these classifiers are then employed as preference criteria during the application phase for discerning between antecedent candidates that fulfill all tight conditions. Further details will be given in the full version of the paper.

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 (cf. (Stuckardt, 2001)). Formal specifications of these algorithms will be given in the full version of the paper. ROSANA-ML handles a broad range of entity-specifying expressions, in particular ordinary, possessive, reflexive/reciprocal, and relative pronouns, definite NPs, and names. However, the machine-learning experiments carried out in this paper will focus on the important case of third person non-possessive and possessive *pronominal* anaphora.

The ROSANA-ML System has been implemented in Common Lisp. For the task of learning decision tree classifiers from the training data, the C4.5 implementation for Unix of the University of Regina[4] is employed.

4. Basic Layout of Experiments

A series of experiments at different levels of consideration will be carried out: (1) variation of the sets of (robustly computable) attributes, i.e. of the signature of the feature vectors from which the classifiers shall be learned; (2) variation of training data generation settings; (3) variation of C4.5 decision tree learning settings (pruning confidence factor CF); (4) internal 10-fold cross-validation and learning curve analysis of the decision tree classifier performance; (5) 6-fold cross-validation of the performance measured at the application (anaphor resolution) level. Details regarding these experimental variations will be given in the full paper.

The training and evaluation of the ROSANA-ML system is performed on a corpus of 66 news agency press releases, comprising 24712 words, 406 third-person non-possessives[5], and 246 third-person possessive pronouns. For the first three experimental stages, the corpus has been firmly partitioned into a training subset (31 documents, 11808 words, 202 non-possessives, 115 possessives) and an evaluation subset (35 documents, 12904 words, 204 non-possessives, 131 possessives); during the cross-validation stages, further partitions are generated randomly.

The anaphor resolution performance is evaluated w.r.t. two evaluation disciplines: **immediate antecedency** (*ia*) and **non-pronominal anchors** (*na*). In the former 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. In the latter discipline, the performance w.r.t. the (application-relevant) selection of *non-pronominal* antecedents is evaluated: precision and recall measures are defined in the same way; however, only non-pronominal antecedent candidates are considered. (For formal definitions and an in-depth discussion of the evaluation measures, cf. (Stuckardt, 2001).) Thus, the anaphor resolution performance is measured according to the precision/recall tradeoffs (P_{ia}, R_{ia}) and (P_{na}, R_{na}).

5. Experiments and Empirical Results

5.1. Finding the Optimal Signature and Settings

In figure 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 contain the evaluation results for the signatures σ_{n0}, σ_{n1}, and, respectively, σ_{full} (first level of experimental variation). Inside these groups, the training data generation settings are variated (second level). At these stages

[4]Release 8, available at `http://www.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html`

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

experiment	antecedents (P_{ia}, R_{ia})		anchors (P_{na}, R_{na})	
	PER3	POS3	PER3	POS3
ROSANA (manually), $[d_1^{31}, d_{32}^{66}]$	$(0.71, 0.71)$	$(0.76, 0.76)$	$(0.68, 0.67))$	$(0.66, 0.66)$
(1) σ_{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) ∧ no cataphors	$(0.62, 0.62)$	$(0.77, 0.77)$	$(0.57, 0.56)$	$(0.70, 0.70)$
(1^{tc}) = (1) ∧ type(α)-specific classifiers	$(0.61, 0.60)$	$(0.69, 0.69)$	$(0.56, 0.55)$	$(0.66, 0.65)$
(1_{nc}^{tc}) = (1^{tc}) ∧ no cataphors	$(0.63, 0.63)$	$(\mathbf{0.76}, \mathbf{0.76})$	$(0.60, 0.59)$	$(\mathbf{0.73}, \mathbf{0.73})$
$(1_{nc}^{tc}+)$ = (1_{nc}^{tc}) ∧ no recency filter	$(0.63, 0.63)$	$(0.73, 0.73)$	$(0.58, 0.58)$	$(0.63, 0.63)$
(2) σ_{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) ∧ no recency filter	$(0.60, 0.60)$	$(0.70, 0.70)$	$(0.52, 0.50)$	$(0.61, 0.60)$
(2_{nc}) = (2) ∧ no cataphors	$(0.63, 0.62)$	$(0.74, 0.74)$	$(0.57, 0.57)$	$(0.66, 0.66)$
(2^{tc}) = (2) ∧ type(α)-specific classifiers	$(0.60, 0.60)$	$(0.70, 0.70)$	$(0.56, 0.55)$	$(0.68, 0.67)$
(2_{nc}^{tc}) = (2^{tc}) ∧ no cataphors	$(0.63, 0.63)$	$(0.73, 0.73)$	$(0.60, 0.59)$	$(0.65, 0.65)$
(3) σ_{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) ∧ type(α)-specific classifiers	$(0.61, 0.61)$	$(0.69, 0.69)$	$(0.57, 0.56)$	$(0.63, 0.62)$
(3_{nc}^{tc}) = (3^{tc}) ∧ no cataphors	$(0.62, 0.62)$	$(0.75, 0.75)$	$(0.57, 0.56)$	$(0.64, 0.64)$
(3+) = (3) ∧ no recency filter	$(0.60, 0.59)$	$(0.69, 0.69)$	$(0.49, 0.49)$	$(0.57, 0.57)$
$(3^{tc}+)$ = (3+) ∧ type(α)-specific classifiers	$(0.62, 0.61)$	$(0.68, 0.68)$	$(0.54, 0.53)$	$(0.64, 0.63)$
$(3_{nc}^{tc}+)$ = $(3^{tc}+)$ ∧ no cataphors	$(0.62, 0.62)$	$(0.76, 0.76)$	$(0.58, 0.57)$	$(0.68, 0.68)$
$(1_{nc}^{tc}$,15) = (1_{nc}^{tc}) ∧ 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}) ∧ CF=37%	$(\mathbf{0.65}, \mathbf{0.64})$	$(0.72, 0.72)$	$(\mathbf{0.61}, \mathbf{0.61})$	$(0.64, 0.64)$
$(1_{nc}^{tc}$,50) = (1_{nc}^{tc}) ∧ 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}) ∧ 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}) ∧ CF=75%	$(0.62, 0.62)$	$(0.72, 0.72)$	$(0.56, 0.56)$	$(0.61, 0.61)$
$\mathbf{(1_{nc}^{tc},h)}$ = (1_{nc}^{tc}) ∧ CF=$\binom{37}{25}$%	$(\mathbf{0.65}, \mathbf{0.64})$	$(\mathbf{0.76}, \mathbf{0.76})$	$(\mathbf{0.62}, \mathbf{0.61})$	$(\mathbf{0.73}, \mathbf{0.73})$

Figure 2: evaluation results

of experimentation, the partition of the corpus into training data and evaluation data remains fixed ($[d_1^{31}, d_{32}^{66}]$).

Regarding 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. More importantly, an in-depth qualitative analysis of the failure cases concerning the selection of immediate antecedents revealed that a substantial amount of incorrect decisions are due to the fact that, in its initial version, ROSANA-ML was not biased against cataphora (i.e. cases of anaphora with antecedents surface-topologically following the anaphor), and, moreover, ROSANA-ML failed to learn the respective (knowingly *globally* useful) preference from the training data since the cospecification information employed at the learning-relevant level of *individual* (local) decisions is inherently symmetrical.[6] This observation gave rise to a further variation at the level of feature vector generation settings: *eliminating instances of cataphoric resumption* in the training as well as the application case.

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}.

[6] In ROSANA, this negative preference was handcoded in the cataphora malus factor applied in the antecedent scoring phase.

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

Figure 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 series of cases [(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 latter case, in which the *no cataphor* as well as the *type-specific classifier* settings are activated, there is a slight gain w.r.t. immediate antecedency of possessive pronouns, and a slight to considerable gain concerning the *nonpronominal anchor*s scores for non-possessives and possessives. A possible explanation might be given by referring to the respective training set sizes, which are displayed in figure 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 might thus be explained as follows: if the amount of available data is sufficiently large, the adulterating effect of artificially enlarging the training set prevails; if, however, learning data is sparse, the overall effect might

be positive .
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^{tc}_{nc}) experiment, which, in terms of overall (averaged) performance, may be considered as the empirically best constellation: 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 clearly prevail. In this case, the advantage of employing specialized classifiers might outweigh the disadvantage of the small training set size since the number of attributes of signature σ_{n0} is considerably lower than in the case of σ_{full} (14 vs. 38).
Thus, the settings of the experiment (1^{tc}_{nc}) have been taken as the starting point of further variations at the level of the C4.5 decision tree learning proper, 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 the fourth group of rows in figure 2, the original setting of CF=25% seems to yield the best scores; classifiers for non-possessives, however, should be determined with a slightly higher CF of 37%. Again, a possible explanation might be given by referring to the different training set sizes: for non-possessives, more training cases are available, resulting in a decision tree that better generalizes, thus allowing for a lower amount of pruning, i.e. a higher pruning confidence factor.
The row (1^{tc}_{nc},h) displays the evaluation results of a "hybrid" setting in which specialized classifiers for non-possessives and possessives are computed with the respectively "best" choices of CF factor values.

5.2. Internal Cross-Validation and Learning Curves

According to the results of the internal cross-validation of the PER3 and the POS3 decision trees, the performance w.r.t. the classification of NON_COSPEC instances lies above 96%, whereas COSPEC instances are recognized with an accuracy of approximately 60%. An analysis of the learning curve reveals that the training corpus should comprise at least 5,000 vectors (1,000 COSPEC instances, 4,000 NON_COSPEC instances) for obtaining a performance near the empirically observed optimum of 60% regarding COSPEC cases.
At first sight, the low accuracy obtained for the COSPEC cases seems to impose a problem. Regarding the application task of anaphor resolution and the ROSANA-ML algorithm, however, it is of primary importance not to misclassifiy the NON_COSPEC cases; erroneously classifying instances of the COSPEC class as NON_COSPEC members is only problematic if there are no further correct antecedent candidates which are correctly classified.
In the full version of the paper, comprehensive cross-validation results will be given in the form of confusion matrices for the PER3 and POS3 classifiers; moreover, plots of the learning curves will be included.

5.3. Application-Oriented Cross-Validation

The results of the 6-fold cross-validation at the application (anaphor resolution) level are displayed in figure 4. The data has 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. $[d_1^{66} \setminus d_{si}, d_{si}], \ldots, 1 \leq i \leq 6$.
Regarding the results of the six base experiments, the variance is considerable. Similar observations have already been made during the evaluation of the manually designed ROSANA system. Thus, the variance seems to be determined by the individual "difficulty" of the document sets w.r.t. the anaphor resolution task rather than being an indicator of a specific problem of the machine-learning approach. 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^{tc}_{nc},h) experiment. Overall, the results of the application-oriented cross-validation can be interpreted as confirming the figures obtained in the original experiment (1^{tc}_{nc},h).

5.4. Trading off Recall for Precision

Based on the above identified empirically optimal settings (1^{tc}_{nc},h), a series of further experiments has been carried out that addresses the question whether, by looking at the additional quantitative data given at the leaves of the C4.5 decision trees, it is possible to gradually bias ROSANA-ML towards **high-precision anaphor resolution**. Besides the category prediction, the decision tree leaves contain further information regarding the number μ of *matching instances*, and the number ε of *missclassified instances* of the training data. By computing the quotient $\frac{\varepsilon}{\mu}$, it should thus be possible to derive a coarse estimate of the classification error probability of the particular leave.
This information may now be used to gradually bias ROSANA-ML towards high-precision anaphor resolution. The base version of the algorithm 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 $\frac{\varepsilon}{\mu}$, this criterion might be refined as follows: prefer COSPEC candidates over NON_COSPEC candidates; at the secondary level, prefer COSPEC candidates with smaller error estimate $\frac{\varepsilon}{\mu}$ over COSPEC candidates with higher $\frac{\varepsilon}{\mu}$, and prefer NON_COSPEC candidates with higher error estimate $\frac{\varepsilon}{\mu}$ over NON_COSPEC candidates with lower $\frac{\varepsilon}{\mu}$. By imposing a minimum threshold θ on this preference ordering, i.e. by eliminating all candidates that fall below the threshold θ, a bias might be imposed that gradually trades off recall for precision.
In figure 5, the results of four respective experiments are displayed. (Full details regarding the experimental settings will be given in the final version of the paper.) The scores obtained for the different evaluation disciplines indicate that, indeed, by employing the quantitative data of the decision tree leaves in the above-described way, one obtains a suitable means for gradually biasing ROSANA-ML towards high precision.

experiment	antecedents (P_{ia}, R_{ia})		anchors (P_{na}, R_{na})	
	PER3	POS3	PER3	POS3
(1^{tc}_{nc},h) $[d_1^{31}, d_{32}^{66}]$, cf. figure 2	$(0.65, 0.64)$	$(0.76, 0.76)$	$(0.62, 0.61)$	$(0.73, 0.73)$
(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	$(0.66, 0.66)$	$(0.75, 0.75)$	$(0.62, 0.62)$	$(0.68, 0.68)$

Figure 4: 6-fold cross validation of anaphor resolution results

experiment	antecedents (P_{ia}, R_{ia})		anchors (P_{na}, R_{na})	
	PER3	POS3	PER3	POS3
(1^{tc}_{nc},h) $[d_1^{31}, d_{32}^{66}]$, cf. figure 2	$(0.65, 0.64)$	$(0.76, 0.76)$	$(0.62, 0.61)$	$(0.73, 0.73)$
(1^{tc}_{nc},h,*p*) = (1^{tc}_{nc},h) $\wedge$ θ = (1.0,1.0)	$(0.79, 0.51)$	$(0.86, 0.60)$	$(0.75, 0.45)$	$(0.83, 0.54)$
(1^{tc}_{nc},h,p^-) = (1^{tc}_{nc},h) $\wedge$ θ = (1.0,0.25)	$(0.74, 0.56)$	$(0.78, 0.63)$	$(0.71, 0.52)$	$(0.76, 0.59)$
(1^{tc}_{nc},h,p^+) = (1^{tc}_{nc},h) $\wedge$ θ = (0.25,1.0)	$(0.81, 0.45)$	$(0.89, 0.50)$	$(0.74, 0.36)$	$(0.67, 0.30)$
(1^{tc}_{nc},h,p^{++}) = (1^{tc}_{nc},h) $\wedge$ θ = (0.1,1.0)	$(0.83, 0.31)$	$(1.00, 0.17)$	$(0.80, 0.08)$	$(1.00, 0.12)$

Figure 5: evaluation results: trading off recall for precision

6. Comparison

6.1. ROSANA-ML vs. ROSANA

The comparison of the evaluation results for experiment (1^{tc}_{nc},h) with the scores of the manually designed ROSANA system on $[d_1^{31}, d_{32}^{66}]$ (cf. figure 2) leads to a nonuniform assessment. Whereas ROSANA-ML performed better w.r.t. nonpronominal anchor determination for possessive pronouns, the results for non-possessives definitely deteriorated. At first sight, this seems to be surprising since, as observed in section 5.2., regarding the classifier for possessives, the training set size of 2670 is clearly too small to arrive at the possible accuracy level of around 60% w.r.t. the recognition of COSPEC cases. However, it has to be kept in mind that, according to the results that have been determined for the manually designed ROSANA system, possessives generally seem to be easier to resolve than non-possessives; the evaluation scores of ROSANA-ML can be interpreted as a further support of these findings.

Hence, if one takes into account the efforts that went into the refinement of the preference factors employed in manually designed systems, the results can be seen 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.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 5.2., at least in the application-oriented cross-validation experiments, the training set size should have been sufficiently large ($> 8,000$) to arrive at the possible accuracy level of around 60% w.r.t. 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 is available.

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

In their machine-learning approach to anaphor resolution of Japanese texts, Aone and 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. A first, coarse impression, however, might 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 recall measures are employed,[7] these results may be compared to the scores of the high-precision anaphor resolution experiments that are summarized in table 5. 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 and Bennett (1995) approach outperforms ROSANA-ML, evidence clearly is the other way around if one takes the zero pronoun figures (0.76,0.38) as the base of comparison. As pointed out by Aone and Bennet, however, quasi-zero pronouns are more easily to resolve since, by definition, they always cospecify with a local subject, and, hence, they might be interpreted by purely syntactical means. Consequently, the zero pronoun scores may be regarded as the more suitable reference for comparision, thus urging upon the conclusion that the methodology of ROSANA-ML of applying machine-learning techniques to derive anaphor resolution *preference* strategies can be regarded to be superior to the unfocused learning approach employed by Aone

[7] 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 ones employed in the paper at hand.

and 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 that match one of these rules; all other decisions are left open. 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 careful manual design of a high-precision ruleset outperforms the machine-learning-based high-precision approach, which was obtained above as a side-product by referring to quantitative information available at the decision tree leaves. It remains to be seen whether better precision-biases might be obtained when employing decision trees which have been learned over larger amounts of training data and which, hence, are expected to provide more reliable data suitable for computing, according to the method outlined in 5.4., better estimates of the classification error probability.

7. Conclusion and Further Research

Overall, the results determined for ROSANA-ML can be regarded to be promising. According to the results of the above experiments, it can be concluded that, by employing a machine-learning-based approach to anaphor resolution, results that at least compare with those of the best manually tuned systems can be reached. Moreover, the investigation has given first evidence that, by biasing ROSANA-ML towards precision, better (P,R) tradeoffs may be obtained than those identified by Aone and Bennett (1995).

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 necessitate that the set of features over which the classifiers are learned is suitably supplemented. Finding suitable candidate features, however, can be considered to be a hard and time-consuming intellectual task, thus illustrating that even machine-learning approaches do not in any case free the knowledge engineer from manual fine-tuning. In this particular case, the task might be immediately compared with the manual choice 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 almost certainly enhance the need for bigger training corpora. 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-specifical assignment of preference strategies, a goal that has been put forward as a consequence of the evaluation of the manually designed ROSANA system.

Finally, based on these further experiments on larger and heterogenous 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? The qualitative analysis of the learned classifiers might ultimatively shed new light on the empirical foundation of classical strategies for determining salience, including theories of attentional focussing such as centering.

8. 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.

Chinatsu Aone and Scott William Bennett. 1996. Applying machine learning to anaphora resolution. In S. Wermter, E. Riloff, and G. Scheler, editors, *Connectionist, statistical and symbolic approaches to learning for Natural Language Processing*, pages 302–314, Berlin. Springer.

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)*.

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.

Ruslan Mitkov. 2001. Outstanding issues in anaphora resolution. In Alexander Gelbukh, editor, *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Mexico City*, pages 110–125.

Roland Stuckardt. 1997. Resolving anaphoric references on deficient syntactic descriptions. 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 30–37.

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