In: Proc. 2003 Int. Symp. Reference Resolution and Its Application to QA and TS (ARQAS), Venice, June 2003, 33-41.

Coreference-Based Summarization and Question Answering: a Case for High Precision Anaphor Resolution

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

Abstract

Approaches to Text Summarization and Question Answering are known to benefit from the availability of coreference information. Based on an analysis of its contributions, a more detailed look at coreference processing for these applications will be proposed: it should be considered as a task of anaphor resolution rather than coreference resolution. It will be further argued that high precision approaches to anaphor resolution optimally match the specific requirements. Three such approaches will be described and empirically evaluated, and the implications for Text Summarization and Question Answering will be discussed.

1 Introduction

Text Summarization (TS) and Question Answering (QA) are generic applications that highly benefit from the availability of an enhanced software technology for the content-oriented analysis of potentially noisy textual data. Recent research has shown that these applications particularly profit from the availability of robust, knowledge-poor solutions to the coreference resolution task as defined at the Message Understanding Conferences MUC-6 and MUC-7.[1] Various research projects have investigated how coreference information can be employed to determine the topics of a text (as relevant for TS), or to determine the contexts that contribute potentially relevant information about entities mentioned in a user query (as relevant for QA).[2]

Beside the work on coreference resolution as fostered by the MUCs, a line of related research addresses the problem of anaphor resolution.[3] These problems are in so far closely related as solutions to the second-mentioned task contribute to performing the first-mentioned task. Importantly, however, they differ with respect to the perspective from which the coreference processing issue is discussed, which is reflected by the different methods that are employed to evaluate software technology for the two tasks: regarding coreference resolution, the scoring procedure refers to the *classes* of coreferring linguistic expressions, whereas the anaphor resolution output is typically assessed by determining the accuracy with which the *antecedent selection* for certain types of anaphoric expressions is performed.

In this paper, the role of coreference information and coreference processing for TS and QA will be explored. Beginning with a brief survey of recent work, the potential contributions of coreference information to QA and TS are identified. Based on the identification of different ways of employing coreference information for these applications, it will be studied in detail which kind of coreference processing is needed, and which requirements an algorithm should meet in order to optimally support the solution of these tasks. This involves a theoretical analysis as well as a look at empirical data. According to

[1] cf. Hirschman (1998)

[2] e.g., cf. Baldwin and Morton (1998), Azzam, Humphreys, and Gaizauskas (1999), Breck et al (1999), Morton (1999)

[3] cf. the monograph of Mitkov (2002)

the results of these investigations, coreference processing for TS and QA should be looked at in more detail: it should be considered as a task of anaphor resolution rather than coreference resolution. Moreover, approaches to anaphor resolution should be biased towards high precision in order to match the specific requirements. Three such approaches will be described and empirically evaluated, and the implications for TS and QA will be discussed.

2 Coreference Information for Text Summarization and Question Answering

Various research projects have explored coreference-based approaches to TS and QA. A brief survey of some representative approaches will be given.

2.1 Text Summarization

Baldwin and Morton (1998) investigated coreference-based TS in an information retrieval (IR) scenario in which automatically generated document summaries are used to support relevance judgments of the IR user. Basically, coreference analysis is employed in two processing stages: (1) retrieving referential relations between the terms of the original IR query and the terms of the documents that are considered, with respect to the query, to be of highest relevance by the IR engine; (2) generation of the document summary by identifying the sentences in which entities of the query that have been identified at stage (1) occur. For solving the second-mentioned problem, the system follows the coreference chains and heuristically selects a subsequence of sentences that, according to further criteria, are judged to be of highest relevance; at this stage, the approach further takes care to provide lexically informative substitute expressions for anaphors (in particular pronouns) that may, out of their original context, become incomprehensible.[4]

The approach of Azzam, Humphreys, and Gaizauskas (1999), too, employs coreference resolution for deriving text summaries. They considered the scenario of *generic* summarization, in which there is no user query that prescribes relevant entities on which the summary should focus. Their algorithm tries to identify a *single* coreference chain pertaining to the central entity the text is about. They further investigated the contribution of a focus mechanism to identify the subsequence of sentences in which this entity is salient.[5]

2.2 Question Answering

QA, too, benefits from the availability of coreference information since it renders possible the identification of contexts in which information regarding the entities a question is about is contributed. A formal definition of a QA scenario has been provided and investigated at the TREC evaluation conferences. According to Breck et al (1999) and Morton (1999), whose systems participated at TREC-8, this problem, too, can be solved by employing coreference information in the two stages of (1) relating entities mentioned in the query to the retrieved documents, and (2) looking at the relevant coreference classes and searching the contexts in which these entities occur for information that may contribute to answer the question.[6] As regarding TS, the coreference information is further employed to supply maximally informative substitutes for anaphoric realizations of entities mentioned in contexts that contribute to answering the question.

2.3 Contribution of Coreference to TS and QA

According to the above survey of some representative approaches, the tasks of TS and QA are closely related. Coreference information is employed in different processing stages. For QA and user-focused TS, the first stage consists in relating some terms of the query to coreferring occurrences in the document pool over which the application runs; this may be considered as a special case of the cross-document coreference resolution problem that has been investigated elsewhere.[7] At the second processing stage, three different cases of using exclusively document-local coreference information may be distinguished:

[4] According to Baldwin and Morton (1998), their approach deals with object coreference and event coreference. They further consider the issue of referential relations beyond the identity relation; this, however, merely seems to cover a few domain-specific special cases.

[5] A plethora of further approaches to automatic text summarization has been investigated (cf., e.g., (Mani, 2002)). Recent research has in particular been fostered by the TIPSTER SUMMAC evaluation exercise, cf. (Mani et al., 1998).

[6] An analysis of the type of the expected answer typically supports this process.

[7] cf., e.g., (Bagga and Baldwin, 1998; Ravin and Kazi, 1999)

1. looking at the coreference *classes* of relevant entities in order to retrieve contexts that contribute information potentially relevant for QA;

2. following a coreference *chain* in order to select a sub*sequence* of sentences that constitute, or contribute to, a document summary;

3. identifying coreferring *antecedents for anaphoric occurrences* in order to provide maximally informative substitute expressions (for QA as well as TS).

According to case 1, solutions to the QA task refer to unordered classes, i.e. sets of coreferring occurrences. This seems to indicate that document-local coreference resolution for TS and QA should be addressed by an approach with high empirical performance in the coreference task as formally defined for MUC-6 and MUC-7.[8] The other two cases, however, emphasize asymmetric aspects of coreference, viz. (surface-topologically ordered) chains of coreferring occurrences, or certain types of anaphoric expressions to be substituted by non-anaphoric antecedent expressions. Moreover, as illustrated by figure 1, TS and QA typically employ lexical information in order to identify relevant coreference classes: *lexically informative* occurrences are the typical points of access. This indicates that looking at the cofererence class level only as done by the MUC scoring scheme of Vilain et al. (1996) falls short of capturing certain aspects that are crucial with regard to the applications TS and QA.
It will now be shown that these requirements can be complied with by considering the document-local coreference processing task as a problem of anaphor resolution rather than reference resolution.

3 Coreference processing for TS and QA

3.1 Towards anaphor resolution

Choosing a coreference processing module that optimally supports TS and QA requires appropriate evaluation measures that are expressive with respect to the type of performance that, according to section 2.3, is essential for these applications. To discuss

[8] cf. (Hirschman, 1998) and the respective coreference-class-oriented scoring scheme of Vilain et al. (1996)

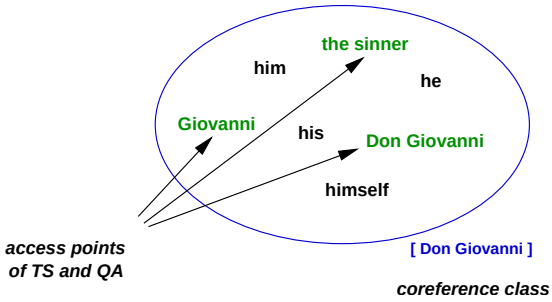

Figure 1: accessing a coreference class via lexically informative occurrences

this issue, an example that illustrates how coreference resolution errors are counted by different evaluation measures proves to be helpful. Figure 2 shows the typical case of a coreference processing error as generated by an employed anaphor resolution system.[9] The coreference classes specified by the key are represented by the dashed boxes. The anaphor resolution system response consists of a set of instances of anaphoric resumption that are represented by arrows pointing from the anaphor to the resumed antecedent; the respective response coreference classes are obtained by computing the reflexive-transitive closure over the individual resumptions and determining the equivalence classes of it. In the configuration shown in figure 2, to an occurrence *he* that belongs to the key coreference class *[Don Giovanni]* an incorrect antecedent has been assigned, which belongs to key class *[Leporello]*.
If the task to be accomplished is considered a problem of coreference resolution the goal of which consists in the computation of the *classes* of coreferring occurrences, the model-theoretic scoring scheme of Vilain et al. (1996) may be employed. According to this scheme, in the configuration of figure 2, there is a single *recall error*, since there is one subclass of the *[Don Giovanni]* key class (represented by the dotted box) that is not connected to the rest of the key class. There is also one single *precision error*: one response class contains a subclass (again, the occurrences in the dotted box) that, according to the key, should have been kept apart from the other occurrences, which belong to class *[Leporello]*.
This scheme, however, does not take into account

[9] As stated above, it can be assumed that even approaches to the coreference resolution task refer to the output of an anaphor resolution stage in order to compute the coreference classes.

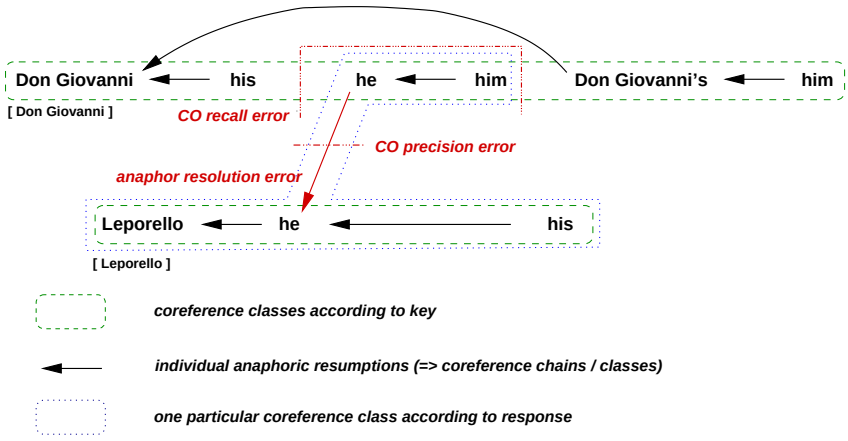

Figure 2: coreference processing error from the perspectives of anaphor vs. coreference resolution

the crucial issues that have been pointed out in section 2.3. Essentially, it considers the following two cases as equal (the arrows represent antecedent choices, "+" denotes "*correct*", "−" stands for "*wrong*"):

(1) Leporello $\overset{-}{\leftarrow}$ *he* $\overset{+}{\leftarrow}$ *him* $\overset{+}{\leftarrow}$ *his*

(2) Leporello $\overset{+}{\leftarrow}$ *he* $\overset{+}{\leftarrow}$ *him* $\overset{-}{\leftarrow}$ *his*

Regarding TS and QA, however, case (1) should be considered worse than case (2). As discussed above and illustrated in figure 1, TS and QA typically access coreference classes or chains via occurrences that are lexically informative. Hence, for these applications, in case (1), only one out of four coreferring occurrences would be found, whereas in case (2), three out of four coreferring occurrences would be retrieved. A similar argument holds with respect to the subtask of providing maximally informative substitute expressions for anaphors in the output of TS and QA. Things are even worse since there is focus-theoretic as well as empirical evidence that the problem of identifying a content-carrying non-pronominal antecedent is considerably harder than identifying an arbitrary antecedent.[10] This implies that the model-theoretic scoring scheme (Vilain et al., 1996) yields results that are, in general, not sufficiently expressive with respect to the contribution of coreference resolution to TS and QA.

The refinement of model-theoretic coreference scoring that is suggested by Bagga and Baldwin (1998) (B-CUBED scoring algorithm) weights errors by taking into account the number of occurrences of the affected class and the relative sizes of the induced subclasses. It meets the specific requirements of evaluating cross-document coreference resolution systems, whereas it does not comply with the above identified requirements.

To achieve the required sensitivity, the problem should be looked at in more detail. Coreference processing for TS and QA should be considered as a task of anaphor resolution rather than coreference resolution, and one should depart from evaluation schemes merely grounded on coreference classes. Formal measures for the evaluation of anaphor resolution systems should be employed. Let (α, γ) be a pair consisting of an anaphoric occurrence α and an antecedent occurrence γ determined by the anaphor resolution system. (If, for α, no antecedent has been determined, then γ is empty.) The scoring is based on a disjoint partition of the pairs (α, γ) output by the anaphor resolution system into the following sets: o_{++} (α and γ corefer), o_{+-} (α and γ do not corefer), $o_{+\_}$ (γ empty, no antecedent assigned), $o_{+?}$ (γ denotes a spurious occurrence). By distinguishing between measures of precision and recall, the second-mentioned of which takes into account the cases with empty antecedent γ as well, one obtains the definitions:[11]

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

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

[10] Pronouns typically refer to focused entities. Hence, they are, with higher probability, correct antecedents (cf. (Stuckardt, 2001)).

[11] For further details cf. (Stuckardt, 2001).

These measures will be employed to determine the performance regarding the identification of arbitrary coreferring antecedents (immediate antecedency *(ia)* discipline). Now, essentially, to take into account the central issue of anchoring pronouns in information-carrying occurrences, which constitute the potential access points of TS and QA, the evaluation discipline of *non-pronominal anchoring (na)* will be considered. It employs the same P and R measures; however, only pronominal anaphors α are taken into consideration, and antecedents γ are required to be lexically informative.[12]

It has to be emphasized that it is not proposed to reduce coreference processing for TS and QA to the mere task of providing lexically informative substitutes for pronouns or to pronominal anaphor resolution. Obviously, *general* coreference information has to be provided in order to comply with the requirements of TS and QA identified in section 2.3. According to the above analysis, however, technology assessment by model-theoretic coreference scoring falls short of adequately capturing the important case of lexically anchoring pronouns, which is a harder problem than coreference class determination according to the MUC task.[13] Moreover, the provision of lexically informative substitutes for pronouns is known to significantly enhance the performance of QA systems, as has been shown by the empirical investigation of Vicedo and Ferrández (2000a).[14] Clearly, the actual contribution of pronominal anaphor resolution depends on the density of pronoun occurrences relevant to the specific QA task; as shown by Vicedo and Ferrández (2000b), it may be moderate in certain cases.

3.2 The case for high prec anaphor resolution

By further reflecting upon how, according to section 2.3, TS and QA employ referential information, additional evidence regarding the specific anaphor resolution strategy that optimally supports these applications can be obtained.

With respect to the coreference chains sought by TS, *recall errors* can be expected to affect the quality of the summarization output only weakly since these errors tend to have local impact only. Typically, regarding a particular coreference chain, as illustrated by figure 2, there will be local sequences of *pronouns* that, due to single anaphor resolution errors, are not connected to the chain. However, subsequent occurrences that carry more referentially discriminative information will be correctly resolved; thus, the interrupted chain will be resumed. This is further supported by empirical data that will be presented below. Since the spread of the chain can thus be expected to still cover the whole document, and since the summary is typically constructed by selecting a *subsequence* of occurrences, the loss should not be too big. *Precision errors*, on the other hand, potentially affect the output quality: if occurrences are erroneously identified as coreferring, the summary may contain irrelevant sentences; the reader may be further misled if incorrect substitute expressions for pronouns are provided.

Regarding the coreference classes sought by QA, *recall errors* do have potential impact since information contributed by a context of a not-found coreferring occurrence gets lost. However, the document pool over which QA is performed may exhibit redundancy and the sought information may be retrieved from elsewhere; in fact, this has been empirically observed by Vicedo and Ferrández (2000b) during the TREC-9 evaluation. *Precision errors* are critical since they can lead to a wrong answer derived from a context of a non-coreferring occurrence, including an incorrect substitute expression.

Thus, precision errors can be expected to cause more potential damage with respect to the applications TS and QA than recall errors. Hence, strategies to high precision anaphor resolution shall be explored.

4 Three approaches to robust high precision anaphor resolution

In order to see which level of performance can be reached, three approaches to high precision anaphor resolution will be investigated. The subsequent dis-

[12] This proposal can be rendered even more generally: the performance with respect to relating *lexically less informative* (typically anaphoric) occurrences to *lexically more informative* occurrences should be measured. Pronouns are the most important and easy-to-recognize special case: they are lexically less informative than any common NP or name occurrence.

[13] A detailed explanation is given in (Stuckardt, 2001).

[14] Vicedo and Ferrández (2000a) focus on the case of QA by text snippet (viz., sentence) extraction. Corresponding to the two different ways of employing document-local coreference information for QA identified in section 2.3, they prove the positive effects of substituting pronouns that refer to entities (1) mentioned in the query, and (2) to be mentioned in the answer.

cussion focuses on the subproblem of third-person pronominal anaphora, the interpretation of which is known to be of particular importance to TS and QA (cf., e.g., (Vicedo and Ferrández, 2000a)).[15] The approaches should work robustly on texts of arbitrary domains, i.e. under the side condition of knowledge-poor processing of potentially noisy data. The robust syntactic salience-based anaphor resolution system ROSANA and its machine-learning-based descendant ROSANA-ML of Stuckardt (2001; 2002) are taken as the starting points.[16]

4.1 ROSANA with CogNIAC high prec ruleset

The first approach consists in the partial reimplementation of the CogNIAC system of Baldwin (1997), which is designed to achieve high precision pronoun resolution. CogNIAC combines the morphological agreement and syntactic disjoint reference filters with six antecedent selection rules, each of which covers one specific situation in which there seems to be little or no ambiguity regarding the antecedent choice. The antecedent filters are employed prior to the six high precision rules, which are applied in order of increasing ambiguity: if a rule applies, the respective candidate will be chosen; if no rule applies, the anaphor remains unresolved.
The CogNIAC system of Baldwin (1997) requires full parses, whereas its reimplementation ROSANA-CogNIAC, which combines the *robust* antecedent filters of ROSANA with the high precision ruleset of CogNIAC, works on partial parses, and, hence, meets the robustness requirements.

4.2 ROSANA with salience threshold

A second approach to high precision pronoun resolution consists in an even more immediate adaptation of the antecedent selection phase of classical, salience-based anaphor resolution algorithms:

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

The rationale behind this strategy is that salience does not only constitute a base for heuristically comparing the relative plausibility of the candidates (and choosing the one with highest salience); in addition, it can be employed as an heuristic estimate of the probability that an individual candidate is a correct antecedent, thus allowing to decline candidates with low salience in order to avoid risky decisions.
By accordingly modifying the antecedent selection step of ROSANA, the system ROSANA-θ is obtained.

4.3 ROSANA-ML towards high precision

Another approach to high precision pronoun resolution has been investigated as part of the research on the machine-learning-based approach ROSANA-ML, which employs C4.5 decision tree classifiers for selecting among antecedent candidates fulfilling the filtering criteria.[17] Basically, the decision trees represent classifier functions which map pairs of anaphors and antecedent candidates (represented as feature vectors) to a prediction $\in \{COREF, NON\_COREF\}$. Beside the primary classification result, the leaves of the decision trees contain additional quantitative information: each leaf provides the total number μ of training cases that match the respective decision path, and the number $\varepsilon \leq \mu$ of these cases that are, through the category prediction of the leaf, wrongly classified. By computing the quotient $\frac{\varepsilon}{\mu}$, it should thus be possible to derive an estimate of the classification error probability of the particular leaf.
This information can be employed to gradually bias ROSANA-ML towards high precision. For this end, the preference criterion of ROSANA-ML, which refers to the (heuristical) decision tree predictions and employs surface-topological distance as the secondary criterion, has been modified by adding a threshold θ that imposes bounds on the admissible classification error probability estimates $\frac{\varepsilon}{\mu}$.[18] This yields the system ROSANA-ML-θ, the degree of inclination towards precision of which depends on θ.

[15] As argued above, since general coreference information is required, pronoun resolution has to be supplemented by strategies dealing with other types of referring expressions, in particular names and common NPs.

[16] ROSANA and ROSANA-ML interpret names and definite NPs as well, and perform general coreference resolution. Only the discussion focuses on pronoun resolution issues.

[17] ROSANA-ML has been trained and thoroughly evaluated (including intrinsic 10-fold and extrinsic 6-fold cross-validation of the learned classifiers) on a corpus of 66 press releases (cf. section 4.4). Full details are given in (Stuckardt, 2002).

[18] Details are given in (Stuckardt, 2002).

experiment	antecedents (P_{ia}, R_{ia})		anchors (P_{na}, R_{na})	
	PER3	POS3	PER3	POS3
(0) ROSANA (salience-based)	$(0.71, 0.71)$	$(0.76, 0.76)$	$(0.68, 0.67)$	$(0.66, 0.66)$
(1) ROSANA-CogNIAC	$(0.66, 0.49)$	$(0.82, 0.53)$	$(0.62, 0.42)$	$(0.79, 0.45)$
(2) ROSANA-CogNIAC, (R6)'	$(0.74, 0.59)$	$(0.82, 0.53)$	$(0.71, 0.53)$	$(0.77, 0.45)$
(3) ROSANA-θ (θ = 90)	$(0.75, 0.67)$	$(0.79, 0.74)$	$(0.74, 0.62)$	$(0.72, 0.63)$
(4) ROSANA-θ (θ = 110)	$(0.79, 0.62)$	$(0.81, 0.50)$	$(0.77, 0.56)$	$(0.74, 0.38)$
(5) ROSANA-ML-θ, p	$(0.79, 0.51)$	$(0.86, 0.60)$	$(0.75, 0.45)$	$(0.83, 0.54)$
(6) ROSANA-ML-θ, p^-	$(0.74, 0.56)$	$(0.78, 0.63)$	$(0.71, 0.52)$	$(0.76, 0.59)$
(7) ROSANA-ML-θ, p^+	$(0.81, 0.45)$	$(0.89, 0.50)$	$(0.74, 0.36)$	$(0.67, 0.30)$
(8) ROSANA-ML-θ, p^{++}	$(0.83, 0.31)$	$(1.00, 0.17)$	$(0.80, 0.08)$	$(1.00, 0.12)$

Figure 3: evaluation results of the high precision approaches on a corpus of *News Agency Press Releases*

4.4 Evaluation

Figure 3 displays evaluation results of the above approaches on a corpus of 35 news agency press releases, comprising 12,904 words, 204 third-person non-possessives, and 131 third-person possessives.[19] In row (0), the results of the original version of ROSANA are shown. Arbitrary immediate antecedents (*ia* task) are chosen with an accuracy (P=R) of 0.71 for non-possessives (PER3), and with an accuracy (P=R) of 0.76 for possessives (POS3). If the more difficult *na* task of identifying nonpronominal antecedents is considered, results deteriorate to $(0.68, 0.67)$ and $(0.66, 0.66)$, respectively. This provides empirical support for the argument of section 3.1 according to which the problem of identifying information-carrying antecedents is considerably harder than the problem of identifying an arbitrary coreferring antecedent.

The subsequent rows display results for the three high precision approaches. ROSANA-CogNIAC's scores are given in rows (1) and (2). Two versions of ROSANA-CogNIAC are considered since it turned out that one of the original rules of CogNIAC (rule 6, dealing with intersentential subject preference) should be modified in order to achieve better results on the Press Releases texts.[20] The cumulated performance with respect to the *ia* discipline (covering third-person non-possessive, possessive, and relative pronouns)[21] amounts to 0.78 precision at 0.60 recall. It thus lags behind original CogNIAC's performance of 0.92 precision at 0.64 recall, which was determined by Baldwin (1997). This might be attributed to the harder conditions of robust processing under which ROSANA-CogNIAC has been run.

ROSANA-θ has been evaluated with a lower and a higher salience threshold. According to the results in rows (3) and (4), precision biasing works; it results in a higher precision at the expense of a lower recall when employing the higher threshold. The same holds with respect to the precision biasing strategy of ROSANA-ML-θ; results for four different threshold settings are displayed in rows (5) to (8).

Obviously, there is no unambiguous winner. Which strategy performs best depends on the targeted (P,R) tradeoff level and on the type of pronoun. E.g., regarding nonpossessives, (4) ROSANA-θ (θ = 110) can be considered to be superior to (2) ROSANA-CogNIAC, (R6)'; regarding possessives, however, empirical evidence is to the contrary. ROSANA-ML-θ (particularly, the p setting) seems to produce the best results on possessives.

According to figure 3, there is a strong correlation between the results in the *ia* discipline and the results in the *na* discipline. Approaches that score high in the first-mentioned discipline typically score high, too, in the second-mentioned discipline. (5) ROSANA-ML-θ, p, e.g., achieves a nonpronominal anchoring performance of 0.83 precision at 0.54 recall which is considerably higher than the figures for (2) ROSANA-CogNIAC, (R6)', which amount to $(0.77, 0.45)$.[22]

ROSANA-CogNIAC and ROSANA-θ have been further evaluated on a corpus of a different genre

[19] For the system development, a separate training corpus of 31 press releases (11,808 words, 202 non-possessives, 115 possessives) has been employed.

[20] The *previous sentence* notion was slightly weakened to cover *intra*sentential candidates that occur in a previous *clause*.

[21] Relative pronouns are not covered by the results shown in figure 3. They are included here for comparison purposes since they are covered, too, by the performance figures of CogNIAC.

[22] Interestingly, it is even higher than the immediate antecedency figures of the second-mentioned approach.

(plot descriptions of Mozart Operas).[23] This has given evidence that the relative performance of the approaches varies across text genres. Regarding nonpossessives, ROSANA-θ is no longer superior to ROSANA-CogNIAC. Moreover, ROSANA-CogNIAC with the original version of rule 6 now clearly outperforms ROSANA-CogNIAC, (R6)'. This might be due to the resemblance of the genre of the Mozart Operas corpus to the genre of the texts on which the original CogNIAC system was run, which were stories about two persons of different gender.

4.5 Implications for TS and QA

According to section 3.1, since lexically informative occurrences constitute the access points of TS and QA to coreference chains and classes, the discipline of nonpronominal anchoring will be considered here. As displayed in figure 3, regarding nonpossessive pronouns, one can achieve a precision of 0.77 at a recall rate of 0.56 ((4) ROSANA-θ (θ = 110)); compared to the non-biased system ((0) ROSANA (salience-based)), this amounts to a gain of 9% precision at the expense of 11% recall. Regarding possessives, by employing the approach (5) ROSANA-ML-θ, p, a precision of 0.83 at 0.54 recall is reached, which means a gain of 17% precision at the expense of 12% recall.

Concerning TS and QA, this implies that one could expect to reduce the amount of wrongly anchored pronouns from about 33% to about 20% while still retrieving more than 50% of the pronoun occurrences. A further in-depth study has provided empirical support for a specific argument of section 3.2: while high precision anaphor resolution strategies provide an effective means to avoid wrongly anchored subchains of pronouns, they typically do not affect the overall spread of a coference chain. Specifically, the outputs of the non-biased system (0) ROSANA (salience-based) and the high precision approach (1) ROSANA-CogNIAC on the Mozart Operas corpus have been compared. The analysis shows that the spread of the two systems' result coreference chains with respect to the 5 biggest coreference classes[24] of each text is nearly identical.[25] A study of the coreference classes generated by ROSANA-CogNIAC reveals that incorrect antecedent choices are avoided in case of 12 third-person pronouns, which, due to chaining effects as described above, results in a total of 25 third-person pronouns that are no longer anchored to an incorrect lexically informative antecedent. Hence, the high precision strategy can be expected to enhance the quality of the TS output.

Regarding QA, as indicated by the empirical results by Vicedo and Ferrández (2000b), much depends on the document pool over which the application runs (cf. section 3.2). If it exhibits redundancy, it may be reasonable to employ an anaphor resolution strategy with a high degree of inclination towards precision; otherwise, a lower precision bias may yield best results. QA thus seems to be best supported by the threshold-based approaches, which render possible different degrees of biasing. This issue should be studied further by performing respective extrinsic (application-level) evaluation runs.

5 Conclusion and further research

Because of the specific requirements, coreference processing for TS and QA should be looked at in more detail: it should be considered as a task of anaphor resolution rather than coreference resolution. To support the choice of the most appropriate approach, formal evaluation should employ anaphor resolution evaluation measures; in particular, the performance regarding the determination of lexically informative anchors for pronouns should be assessed. In order to optimally contribute to TS and QA, solutions to anaphor resolution should be inclined towards high precision. Three approaches have been investigated. According to formal evaluation, these approaches successfully reduce the amount of wrongly anchored pronouns, while still yielding coreference chains that spread the document as required by TS. QA is expected to benefit from threshold-based approaches, which render pos-

[23] Evaluation figures for these experiments are not included. Since the corpus is quite small, it proved to be impossible to evaluate ROSANA-ML-θ on it, since this approach requires a reasonable amount of training data.

[24] as marked up in an intellectually gathered key

[25] There is a single case in which ROSANA-CogNIAC performs worse, which, however, proved to be not attributable to to the high precision strategy proper. Interestingly, in another case, due to complex processing interdependencies, ROSANA-CogNIAC generated a coreference chain with higher coverage.

sible different degrees of precision bias.

Further research should address the contribution of high precision anaphor resolution at the application (TS, QA) level; with respect to QA, this amounts to continuing the empirical work of Vicedo and Ferrández (2000a), who do not provide a detailed analysis of the impact of pronoun interpretation errors. Regarding the high precision strategies, the issue of genre dependency should be paid attention to. Moreover, the contributions of high precision strategies to sequenced models of anaphor resolution, which employ a series of competence modules of increasing complexity, should be investigated.

Acknowledgements

Thanks to the anonymous reviewer number two for providing a valuable hint at the empirical investigations by Vicedo and Ferrández (2000a; 2000b) on the importance of pronominal anaphor resolution to QA.

References

Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. 1999. Using coreference chains for text summarization. In *Proceedings of the ACL'99 Workshop on Conference and its Applications, Baltimore*, pages 77–84.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreference using the vector space model. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98/ACL'98), Montreal*, pages 79–85.

Breck Baldwin and Thomas S. Morton. 1998. Dynamic coreference-based summarization. In *Proceedings of the 3rd International Conference on Empirical Methods in Natural Language Processing (EMNLP-3), Granada*, pages 1–6.

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.

Eric Breck, John Burger, Lisa Ferro, David House, Marc Light, and Inderjeet Mani. 1999. A sys called qanda. In *Proceedings of the 8th Text Retrieval Conference (TREC-8), Gaithersburgh*, pages 499–506.

Lynette Hirschman. 1998. Muc-7 coreference task definition, version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. Published online, available (June 7, 2003) at http://www.itl.nist.gov/iaui/894.02/.

Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Leo Obrst, Therese Firmin, Michael Chrzanowski, and Beth Sundheim. 1998. The tipster summac text summarization evaluation, final report. Technical Report MTR 98W0000138, MITRE.

Inderjeet Mani. 2002. *Automatic Summarization*. John Benjamins, Amsterdam/Philadelphia.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Longman, London.

Thomas S. Morton. 1999. Using coreference in question answering. In *Proceedings of the ACL'99 Workshop on Conference and its Applications, Baltimore*, pages 85–89.

Yael Ravin and Zunaid Kazi. 1999. Is hillary rodham clinton the president? disambiguating names across documents. In *Proceedings of the ACL'99 Workshop on Conference and its Applications, Baltimore*.

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

Roland Stuckardt. 2002. Machine-learning-based vs. manually designed approaches to anaphor resolution: the best of two worlds. In *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 02)*, pages 211–216.

José L. Vicedo and Antonio Ferrández. 2000a. Importance of pronominal anaphora resolution in question answering systems. In *Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics (ACL'00), Hongkong*, pages 555–562.

José L. Vicedo and Antonio Ferrández. 2000b. A semantic approach to question answering systems. In *Proceedings of the 9th Text Retrieval Conference (TREC-9), Gaithersburgh*, pages 511–516.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1996. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 45–52. Morgan Kaufmann.