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

Roland Stuckardt

Johann Wolfgang Goethe University
Frankfurt am Main
Germany

roland@stuckardt.de

Text Summarization and Question Answering

- generic applications that require a robust, domain-independent text analysis technology
- **coreference information** is known to be of particular relevance
- **coreference-based Text Summarization (TS):**
  - Baldwin & Morton (1998)
  - Azzam, Humphreys, and Gaizauskas (1999)
  - ...
- **coreference-based Question Answering (QA):**
  - Breck et al. (1999)
  - Morton (1999)
  - ...

Coreference vs. Anaphor Resolution

- **coreference resolution:**

  determine classes of coreferring occurrences (discourse entity mentions)

- **anaphor resolution:**

  assign coreferring antecedents to anaphoric occurrences

- these tasks are **closely related:**
  - solutions to the latter contribute to the former
  - the level of consideration differs

- **claim:**

  coreference processing for TS and QA should be considered as a task of **anaphor** resolution.

Contents

1. Study of coreference-based approaches to TS and QA
2. Analysis of the type of coreference processing needed
3. Conclusions:
   - coreference processing should be considered as a problem of **anaphor** resolution
   - the anaphor resolution engine should be biased towards **high precision**
4. Empirical investigation of three approaches to high precision pronoun resolution
5. Implications

Coreference-Based TS

- **Baldwin & Morton (1998), user-focused TS for IR:**

  stage 1: relating query terms to document terms

  stage 2: exploitation of document-internal coreference:

  (1) selecting important coreference chains

  (2) selecting a subset of important sentences

  (3) supplementing anaphoric expressions with maximally informative expressions

- **Azzam, Humphreys, & Gaizauskas (1999), generic TS:**

  exploitation of document-internal coreference:

  (1) selecting a single important coreference chain

  (2) selecting a subset of important sentences (supported by a focus mechanism)

  (3) supplementing anaphoric expressions with maximally informative expressions

Coreference-Based QA

- **Breck et al. (1999), Morton (1999), TREC-8:**

  stage 1: relating query terms to document terms

  stage 2: exploitation of document-internal coreference:

  (1) searching coreference classes for query-relevant occurrences

  (2) selecting a context that answers the question

  (3) supplementing anaphoric expressions with maximally informative expressions

Use Cases of Coreference for TS and QA

- **two common stages:**

  stage 1: relating query terms to document terms (user-focused TS, QA)

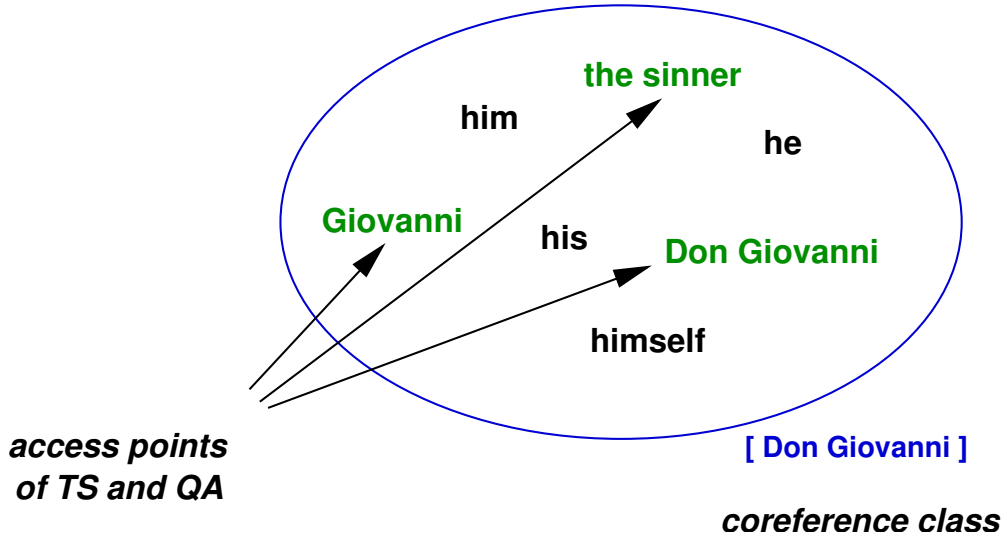
  stage 2: exploitation of document-internal coreference: (TS, QA)

  - QA: looking at coreference **classes** in order to retrieve relevant information
  - TS: traversing coreference **chains** and selecting sub**sequences** of sentences
  - TS and QA: identifying **lexically informative antecedents for anaphors**

$\rightarrow$

- in most cases, an **asymmetric** perspective towards coreference is assumed

Stage 1: Accessing Query-Relevant Coreference Classes

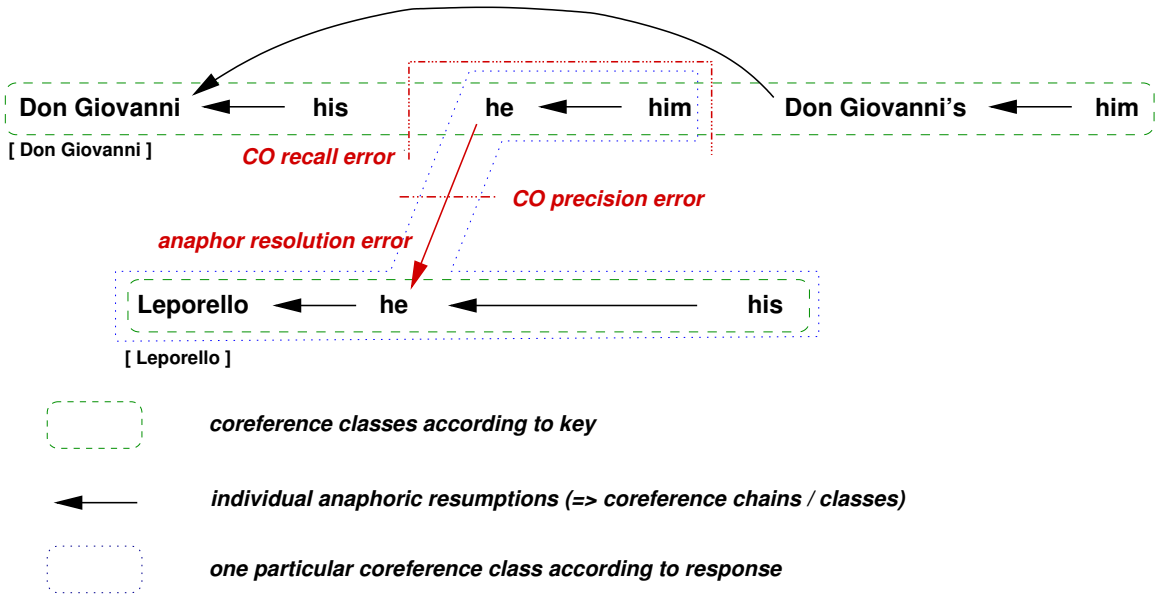


- access via **lexically informative** occurrences
- asymmetry between anaphors (pronouns) and more informative expressions

→

- **how to assess coreference technology for TS & QA?**

Scoring Coreference Interpretation Errors: Coreference vs. Anaphor Resolution



- according to **model-theoretic coreference scoring** (MUC, Vilain et al., 1996), the following errors count equal:

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

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

$\rightarrow$

- **not sufficiently expressive with respect to the contributions to TS & QA**

Towards Scoring Informative Anchors

- let's look at pairs (α, γ) consisting of **anaphors** α and system-determined **antecedents** γ
- disjoint partition of the pairs into the 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)
- precision and recall measures:

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

$\rightarrow$ **two disciplines:**

- immediate antecedency: (P_{ia}, R_{ia})
  γ arbitrary
- **informative (= non-pronominal) anchoring**: (P_{na}, R_{na})
  γ non-pronominal

Interim Result

- of relevance is:

  the anchoring of lexically less informative (typically: anaphoric) occurrences in lexically more informative occurrences

- mere model-theoretic coreference scoring is not sufficiently expressive

- **pronouns** are an important special case:
  - TS: pronouns resume discourse entities in focus
  - QA: significant contribution according to the empirical study by Vicedo & Ferrández (2000a)

- however, it is *not* proposed to reduce coreference processing to mere pronominal anaphor resolution; **general coreference information is required**

The Case for High Prec Anaphor Resolution

- coreference-based TS:
  - precision errors affect the output quality:
    - inclusion of irrelevant sentences
    - incorrect lexically informative expressions
  - recall errors typically have **local** impact only
- coreference-based QA:
  - precision errors affect the output quality:
    - wrong answers
    - incorrect lexically informative expressions
  - recall errors have (possibly limited) impact:
    - relevant contexts may not be found
    - however, the document set may exhibit redundancy (Vicedo & Ferrández (2000b), TREC-9)

$\rightarrow$

- **high precision** anaphor resolution should be investigated

High Precision Anaphor Resolution: Three Approaches

- requirements:
  - domain independency
  - robustness
  - knowledge poorness
- focus on: third-person pronominal anaphora
- starting points:
  - ROSANA, manually designed (Stuckardt, 2001)
  - ROSANA-ML, machine-learning-based (Stuckardt, 2002)

Approach 1: ROSANA-CogNIAC

- based on CogNIAC (Baldwin, 1997)
- covers third-person pronominal anaphora
- **high precision antecedent preference ruleset:**

  (CR1) *unique in discourse*

  (CR2) *reflexive pronouns, nearest possible*

  (CR3) *unique in current and prior*

  (CR4) *possessive pronouns, unique exact match in prior*

  (CR5) *unique in current*

  (CR6) *unique subject in prior (for subject pronouns)*

  otherwise, the pronoun remains *unresolved*

- **new: robust implementation of antecedent filters**
  (in particular, syntactic disjoint reference)

→ **ROSANA-CogNIAC**

Approach 2: ROSANA with Salience Threshold

- immediate adaption of the antecedent selection phase of classical, salience-based approaches:

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

- rationale: salience as an heuristic estimate for
  - the relative plausibility of candidates
  - **the probability that a specific candidate is a correct antecedent**

$\rightarrow$ **ROSANA-θ**

Approach 3: ROSANA-ML towards High Prec

- architecture of ROSANA-ML:
  - antecedent filters are manually designed
  - antecedent preferences are machine-learned (C4.5)
- decision tree lookup predicts CO $\dot{\vee}$ NON_CO
- decision tree lookup yields further information:
  - number μ of **matching** training cases
  - number ε of **wrongly classified** training cases

$\rightarrow$

- estimate $\frac{\varepsilon}{\mu}$ of **classification error probability**
- **candidate acceptance threshold** $\theta := (\theta_{CO}, \theta_{\neg CO})$:
  - accepting CO candidates with $\frac{\varepsilon}{\mu} \leq \theta_{CO}$
  - accepting NON_CO candidates with $\frac{\varepsilon}{\mu} \geq \theta_{\neg CO}$

$\rightarrow$ **ROSANA-ML-θ**

Empirical Experiments, Evaluation Results

- training on 31 news agency press releases, 11,808 words, 202 non-possessives, 115 possessives
- evaluation on 35 news agency press releases, 12,904 words, 204 non-possessives, 131 possessives
- 10-fold / 6-fold cross-validation of ROSANA-ML

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 ($\theta = 90$)	(0.75, 0.67)	(0.79, 0.74)	(0.74, 0.62)	(0.72, 0.63)
(4) ROSANA-θ ($\theta = 110$)	(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)

Findings:

- lexically informative anchoring *(na)* is more difficult than immediate antecedency *(ia)*
- precision biasing works
- **winner** depends on pronoun type and tradeoff level:
  - nonpossessives: ROSANA-θ
  - possessives: ROSANA-ML-θ, p
- ROSANA-CognIAC doesn't reach original CognIAC's performance level ((0.78,0.60) vs. (0.92,0.64)) - presumably due to:
  - conditions of robust processing
  - different genre
- experiments on different corpus indicate **genre dependency**

Implications

- **achievable tradeoffs,** *(na)* **discipline:**
  - nonpossessives: (0.77,0.56) $\doteq$ (+9% P,-11% R)
  - possessives: (0.83,0.54) $\doteq$ (+17% P, -12% R)
- **general interpretation:**
  - reducing pronoun anchoring errors to 20%
  - still retrieving 55% of all pronoun mentions
- **regarding TS**, an in-depth analysis shows:
  - CO chain spread of five biggest CO classes not affected
- **regarding QA**, much depends on
  - the relevance of pronominal occurrences
  - the corpus redundancy

  with respect to the specific task

  $\rightarrow$ presumably best served by threshold-based approaches

Conclusion and Further Research

- coreference processing for TS and QA should be considered as a task of anaphor resolution
- the anchoring of lexically less informative occurrences in lexically more informative occurrences is relevant.
- anaphor resolution should be biased towards high precision
- study of three high precision pronoun resolution approaches:
  - $\approx$ (0.80,0.55) (na) on possessives $\cup$ nonpossessives
  - different tradeoff levels achievable
  - performance depends on genre
  - spread of coreference chains is sustained
- **further reseach:**
  - extrinsic evaluation of high precision anaphor resolution in TS and QA scenarios
  - genre dependency of high precision strategies