

Zur formalen, korpusbasierten Evaluation von Systemen zur Textinhaltsanalyse am Beispiel der Koreferenzresolution

Roland STUCKARDT

1. Einleitung

Die formale Evaluation von Textanalysesystemen blickt auf eine mehr als zehnjährige Geschichte zurück. Zentrale Instanzen mit mittlerweile internationalem Teilnehmerkreis sind die in den USA initiierten Evaluationskonferenzen *TREC (Text REtrieval Conference)*, seit 1992) und *MUC (Message Understanding Conference)*, seit 1986) (vgl. HIRSCHMAN 1998). Während die auf die formale Evaluation klassischer Textretrieval-Systeme abzielenden *TRE*-Konferenzen mittlerweile auch im kontinentaleuropäischen bzw. deutschsprachigen Raum zunehmend ins Zentrum des Interesses rücken, finden die *MUCs* bzw. die in deren Rahmen fokussierte Textanalysetechnologie des *Information Extraction* bislang nur relativ wenig Beachtung. Im Zeitalter der globalen Vernetzung mit weiterhin bestehender Präponderanz textueller Informationsdarbietung muß dieser Sachverhalt verwundern; die zentrale ökonomische Relevanz robuster, massendatentauglicher Softwaresysteme für die Tiefenanalyse von Texten liegt auf der Hand.

Im vorliegenden Aufsatz sollen die wesentlichen Vorteile formaler, korpusbasierter Evaluationen als Motor der Entwicklung von Softwaresystemen des *Information Extraction* rekapituliert und systematisiert sowie zusätzliche Anwendungsgebiete identifiziert werden, die deren Nutzwert erweitern. Im Zentrum der Betrachtungen steht das Textanalyseproblem der *Koreferenzresolution* als eine spezifische Teildisziplin der inhaltlichen Tiefenanalyse, die erstmals im Rahmen der Evaluationskonferenz *MUC-6* als sog. *CO-Task* betrachtet wurde (vgl. VILAIN et al. 1996, GRISHMAN & SUNDHEIM 1996, CoTD 1997). Ausgehend von einer Rekapitulation grundlegender *Basisziele* formaler, korpusbasierter Evaluationen wird eine Reihe

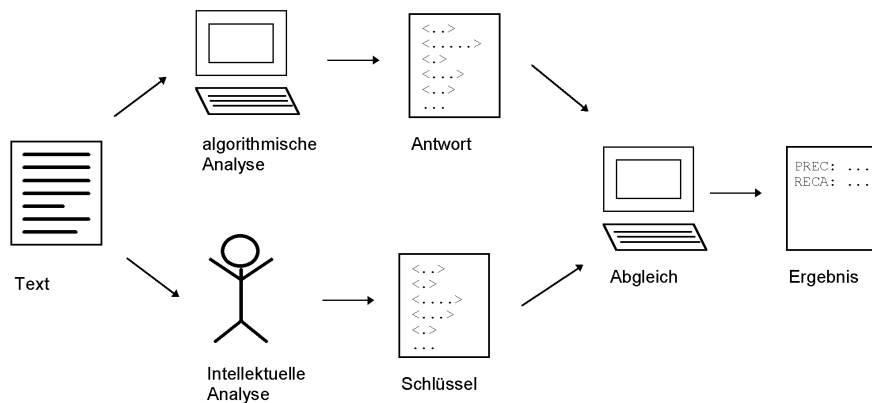


Abbildung 1: Szenario der formalen, korpusbasierten Evaluation

von *Sekundärzielen* identifiziert, die sich aus den Perspektiven von Systementwicklung bzw. Anwendung ergeben. Das Ergebnis besteht in der Definition zweier weiterer Teildisziplinen der Evaluation von Koreferenzresolutions-Systemen. Anhand des Beispiels der Evaluation und Optimierung des Anaphernresolutions-Systems ROSANA (STUCKARDT 1999) wird der praktische Nutzwert des erweiterten Disziplinenkanons demonstriert.

2. Basisziele der formalen, korpusbasierten Evaluation

In Abbildung 1 ist das generische Szenario der formalen, korpusbasierten Evaluation skizziert, das u.a. den *MUCs* zugrunde liegt. Die Vorgehensweise läßt sich folgendermaßen beschreiben:

Für ein anwendungstypisches Textkorpus: Ermittle den Grad der Übereinstimmung zwischen den auf algorithmischem Wege erzeugten Inhaltsbeschreibungen und den auf intellektuellem Wege generierten Inhaltsbeschreibungen.

Vergleichsmaßstab für die systemgenerierten Textinhaltsanalyse-Ergebnisse (*„Antworten“*) sind somit Referenzdaten (*„Schlüssel“*), die durch menschliche Textinterpreten auf intellektuellem Wege erstellt werden.

Am Beispiel des *MUC*-Szenarios zur Evaluation von Systemen zur Koreferenzresolution sollen nun zunächst die Basisziele der formalen, korpusbasierten Evaluation identifiziert werden.

2.1 Basisziele

Zu den grundlegenden Zielen der formalen Evaluation von Textanalysesystemen sind zu zählen:

- Ermittlung numerischer Werte - üblicherweise differenziert in *Precision* und *Recall* -, die die *Qualität der Interpretationsleistung* des jeweiligen Systems aussagekräftig und inhaltsadäquat charakterisieren,
- Herstellung der *Vergleichbarkeit* der Interpretationsqualitäten unterschiedlicher Systeme per Standardisierung von Evaluationsdisziplinen,
- Förderung der Entwicklung tatsächlich *robuster, operationaler* und damit potentiell *anwendungstauglicher Textanalyse-Algorithmen*.

Die Unterscheidung zwischen *Precision* und *Recall* entspricht der in der "Information Retrieval"-Forschung getroffenen¹, wobei in den Definitionen dieser Maße auf Anzahlen retrievalrelevanter Entitäten rekuriert wird, deren Grundmengen in Abhängigkeit der spezifischen Inhaltserschließungsaufgabe festgelegt sind (vgl. Abschnitt 2.2).

Im Mittelpunkt steht zunächst die Evaluation der Leistung des Textanalysesystems per se als *Basistechnologie* ohne Berücksichtigung spezifischer Anforderungen der Entwickler- oder Anwendungssicht.

2.2 Formale Evaluationsmaße für die Koreferenzresolution

Für die Evaluation von Softwaresystemen für das Problem der *Koreferenzresolution* ergibt sich ein geeigneter Ausgangspunkt auf der Grundlage der Beobachtung, daß die auf algorithmischem Wege zu rekonstruierenden Koreferenzbeziehungen zwischen den unterschiedlichen sprachlichen Ausdrücken eines zu interpretierenden Textes als *Äquivalenzklassen einer Äquivalenzrelation* aufzufassen sind. Aus der Perspektive der zu resolvierenden anaphorischen Entitäten existiert demnach i.d.R. eine größere Anzahl als korrekt zu bewertende Vorgänger:

¹ *Precision* = Anzahl korrekter systemseitig generierter geteilt durch die Anzahl aller systemseitig generierten Entitäten, *Recall* = Anzahl korrekter systemseitig generierter geteilt durch die Anzahl gesuchter Entitäten.

- (1) **Gropius traf Behrens, mit dem er die Entwürfe besprechen wollte. Dem Leiter des Bauhauses war es klar, daß er noch viel zu tun bekommen würde.**

Für das zweite pronominale Vorkommen “er” bestehen etwa die Antezedens-Optionen “Leiter des Bauhauses”, “er” und “Gropius”, die allesamt Okkurrenzen ein und desselben Diskursreferenten, d.h. *Mitglieder einer bestimmten Äquivalenzklasse der Koreferenzrelation* sind. Alle genannten Vorkommen sind somit als korrekte Antezedenten der betrachteten Anapher “er” anzusehen. In der *originären Evaluationsdisziplin* für Systeme der Koreferenzresolution ist somit auf *Äquivalenzklassen* von Vorkommen Bezug zu nehmen.

Ausgehend von dieser Beobachtung schlagen Vilain, Burger, Aberdeen, Connolly und Hirschman formale Evaluationsmaße vor, durch die auf die systemseitig generierten bzw. schlüsselseitig vorgegebenen *Äquivalenzklassen* der Koreferenzrelation Bezug genommen wird (vgl. VILAIN et al. 1996). Entsprechend diesem Vorschlag bemißt sich der maximal erzielbare *Recall* je *Schlüssel-Äquivalenzklasse* als die Anzahl der Vorkommen der jeweiligen Klasse minus 1; diese Zahl entspricht der zur Rekonstruktion der jeweiligen Schlüsselklasse minimal notwendigen Anzahl systemseitiger individueller Antezedensentscheidungen. *Recall-Fehler* erschließen sich somit qua Ermittlung der Kardinalität der Partitionierung, die die systemseitig generierte Äquivalenzrelation auf den Schlüsselklassen induziert. In symmetrischer Weise ist die maximal erzielbare *Precision* je *Antwort-Äquivalenzklasse* als die Anzahl der Vorkommen der jeweiligen Klasse minus 1 festgelegt; *Precision-Fehler* schlagen sich in nichttrivialen Partitionierungen nieder, die durch die schlüsselseitig vorgegebene Äquivalenzrelation auf den Äquivalenzklassen der Antwort-Relation induziert werden. Seien RS und RA die Äquivalenzrelationen von Schlüssel bzw. Antwort und $[RS]$ bzw. $[RA]$ die Mengen der entsprechenden Äquivalenzklassen. Seien ferner $\Phi(CA,RS)$ und $\Phi(CS,RA)$ die Äquivalenzrelationen, die sich vermöge Einschränkung der Äquivalenzrelationen RS bzw. RA auf die Elemente der Äquivalenzklassen $CA \in [RA]$ bzw. $CS \in [RS]$ ergeben, und $[\Phi(CA,RS)]$ bzw. $[\Phi(CS,RA)]$ die Mengen der Äquivalenzklassen der eingeschränkten Relationen. Folgende Definitionen formalisieren das zuvor

informell beschriebene Berechnungsverfahren, das den *CO-Task*-Evaluationen von MUC-6 und MUC-7 zugrunde liegt (vgl. STUCKARDT 1999, S. 235ff.):

$$P_{\bar{A}K} = \frac{\sum_{C_A \in [R_A]} (|C_A| - |\Phi(C_A, R_S)|)}{\sum_{C_A \in [R_A]} (|C_A| - 1)} \quad , \quad R_{\bar{A}K} = \frac{\sum_{C_S \in [R_S]} (|C_S| - |\Phi(C_S, R_A)|)}{\sum_{C_S \in [R_S]} (|C_S| - 1)} .$$

2.3 Inhaltliche Zielvorgaben

Von ebenso zentraler Bedeutung sind *inhaltliche Vorgaben*, durch die genauer eingegrenzt wird, welche sprachlichen Ausdrücke als *koreferent* anzusehen sind. Diese nähere Charakterisierung der Aufgabe bildet die Grundlage sowohl für die intellektuelle Erstellung der Schlüsselvorgaben als auch für das Design der Analysealgorithmen der an der Evaluation partizipierenden Systeme. Neben dem Primärziel der *inhaltlichen Adäquatheit* der getroffenen Festlegungen kommt der *intellektuellen Reproduzierbarkeit* eine wichtige Rolle als Prüfstein für die *Gültigkeit* der ermittelten Ergebnisse zu. Die Resultate der Evaluation werden nur dann aussagefähig sein, wenn die (informellen, natürlichsprachigen) Rahmenvorgaben für das Analyseziel von menschlichen Annotatoren hinreichend eindeutig interpretiert werden. Mit anderen Worten: *Die Intercoder-Übereinstimmung* in der Schlüsselerstellung sollte hinreichend groß ausfallen.

Was unter *hinreichender Übereinstimmung* zu verstehen ist, hängt vom Einzelfall, d.h. insbesondere von der Komplexität der jeweils betrachteten Inhaltserschließungsaufgabe ab. In einigen Fällen stehen die Anforderungen der inhaltlichen Adäquatheit und der möglichst guten intellektuellen Reproduzierbarkeit im Widerspruch: Je feinkörniger die Unterscheidung zwischen inhaltlich subtilen Sonderfällen, desto größer die kognitive Komplexität der Aufgabendefinition und somit die Gefahr von Divergenzen in der intellektuellen Interpretation. Ein weiteres Dilemma besteht in bezug auf die Wahl der Annotatoren. Greift man auf *durchschnittliche Sprachbenutzer* zurück, so ist zur Vermittlung der Aufgabendefini-

tion ein i.a. erheblicher Schulungsaufwand zu treiben, um eine hinreichende Intercoder-Reliabilität zu erzielen. Dieser Aufwand ließe sich unter Rückgriff auf linguistisch geschultes Personal reduzieren, das eine höhere Reliabilität auf der Grundlage eines verinnerlichten, evtl. jedoch auch in unerwünschter Weise normierenden theoretischen Vorwissens erzielt. Folglich gibt es gute Gründe für möglichst einfache und klar strukturierte Aufgabenbeschreibungen, die auf der Grundlage intuitiven Sprachverständnisses zugänglich sind.

Wie schwer dieser Anspruch im Einzelfall einlösbar ist, zeigt sich erneut am Beispiel der auf den ersten Blick elementaren Aufgabenstellung der Koreferenzresolution. Die im Rahmen der MUCs entwickelte *Coreference Task Definition* (CoTD 1997) umfaßt 13 Seiten; die auf dieser Grundlage erzielte Intercoder-Übereinstimmung von ca. 81 % (vgl. SUNDHEIM 1996) reicht nicht annäherungsweise an die Zielvorgabe von 95 % (CoTD 1997) heran. Diese Schwierigkeiten sind ein wesentlicher Grund dafür, warum in den MUCs bislang keine komplexeren referentiellen Relationen in die Evaluation einbezogen werden.

3. Sekundärziele

Aus dem Blickwinkel von *Systementwicklung und Optimierung* bzw. aus der *Anwendungsperspektive* lassen sich zusätzliche Anforderungen ableiten, die die Definition weiterer Evaluationsdisziplinen nahelegen.

3.1 Perspektive der Systementwicklung und Optimierung

Im Rahmen der Entwicklung von Systemen der Koreferenzresolution stellt sich das rein inhaltlich motivierte Ziel der Rekonstruktion der Koreferenz-Äquivalenzklassen nicht mehr länger als monolithischer Block dar, sondern es zerfällt in mehrere Teilaufgaben, die auf der Implementierungsebene durch unterschiedliche *Subalgorithmen* gelöst werden. Folglich wären für die Systementwicklerin nähere Informationen betreffend die Performanz der Komponentenmodule bzw. deren spezifische Konfiguration wünschenswert, um gezielte Optimierungen vornehmen zu können.

Textwörter	Schlüssel	System	Schlüssel \cap System
1 Es		+	
2 war			
3 Behrens	+	+	+
4 ,			
5 der	+		
6 gähnte			
7 .			
Σ	2	2	1

Abbildung 2: Beispiel zur Evaluation der Vorkommen-Bestimmung

3.1.1 Evaluation von Subalgorithmen zur Vorkommen-Bestimmung

Ein Teilproblem, dessen Abgrenzung von der eigentlichen Koreferenzresolution aus sowohl implementationstechnischen als auch inhaltlichen Gründen naheliegt, ist die *Bestimmung der referenzierenden sprachlichen Ausdrücke*, die die sog. *Vorkommen* bzw. *Okkurrenzen* konstituieren, über deren Gesamtmenge die Äquivalenzklassen-Partitionierung der originären Aufgabenstellung zu bewerkstelligen ist. Anhand von Beispiel (1) wird deutlich, daß diese Aufgabe nichttrivial ist: Eine einfache Verankerung etwa in den NP-Ausgaben eines robusten Part-of-Speech-Taggers erweist sich im Falle der nichtreferentiellen, expletiven Verwendungsweise des Pronomens “es” als unzureichend. In Systemen der Koreferenzresolution wird das Teilproblem der Vorkommen-Bestimmung i.d.R. durch gesonderte Vorverarbeitungsmodule implementiert, und auch in der in Abschnitt 2.3 diskutierten *Coreference Task Definition* der MUCs bildet die inhaltliche Charakterisierung der Okkurrenzen (bezeichnet als sog. “*markables*”) einen gesonderten Teil der Beschreibung. Zur Unterstützung von Entwicklung und Optimierung der Analysealgorithmen erscheint es demnach sinnvoll, die *Bestimmung der relevanten, referenzierenden Vorkommen als eigenständige Evaluationsdisziplin* zu betrachten. Die originäre, koreferenzklassen-bezogene Evaluation sollte auf die Schnittmenge von schlüsselspezifizierten und systemgenerierten Vorkommen als reduzierte Bezugsmenge der Klassenpartitionierung begrenzt werden; nicht zuletzt werden durch diese konzeptuelle Trennung bestimmte Anomalien in der

Berechnung der klassenbezogenen Evaluationsmaße vermieden (vgl. STUCKARDT 1999, S. 235ff.).

Die Definition geeigneter *Precision*- und *Recall*-Maße gestaltet sich elementar. Relevante Entitäten sind Köpfe von NP, die referenzierende Vorkommen im Sinne der Aufgabenspezifikation induzieren. Der *Precision*-Wert ergibt sich aus dem Quotienten der Anzahl der *korrekten* systemgenerierten Okkurrenzen und der Anzahl *aller* systemgenerierten Okkurrenzen; der *Recall*-Wert ist definiert durch den Quotienten der Anzahl der *korrekten* systemgenerierten Okkurrenzen und der Anzahl *aller korrekten*, d.h. im Schlüssel vorgegebenen Okkurrenzen. Für das Beispiel in Abbildung 2 ergeben sich *Precision*- und *Recall*-Werte von jeweils 0.5.

3.1.2 Optimierung spezifischer Strategien des Kernalgorithmus

Aus der Sicht des Systementwicklers kommt der Evaluation im Rahmen der *Optimierung des Kernalgorithmus* eine weitere wichtige Rolle zu, denn oftmals basieren die einfließenden Rahmenstrategien auf spezifischen Parametern, deren Optimierung noch aussteht. Im speziellen gilt dies für ein in bezug auf das Problem der Koreferenzresolution vorgeschlagenes Analyseparadigma, nach dem die Auswahl des Antezedens aus der Menge potentiell zulässiger Kandidaten unter Rekurs auf einen Set von *Präferenzfaktoren* bewerkstelligt wird. Die relativen Gewichte der Faktoren sowie die anapherntypspezifische Zuordnung adäquater Faktoren-Sets können unter Rekurs auf annotierte Trainingskorpora per schrittweiser Verfeinerung heuristisch optimiert werden. Da die Optimierung anhand eines spezifischen Korpus erfolgt, ist davon auszugehen, daß die getroffenen Festlegungen nicht für alle Texte zu optimalen Auswahlentscheidungen führen. Ein entsprechender Einsatz der korpusbasierten Evaluation erleichtert somit nicht nur die Arbeit des Systemdesigners, sondern liefert darüberhinaus wichtige Anhaltspunkte in bezug auf die zu erwartenden Ergebnisse bzw. die optimale Konfiguration der Rahmenstrategien für Anwendungen auf Textkorpora mit referentieller Kohäsionsstruktur spezifischen Zuschnitts (vgl. Abschnitt 4.2).

Als Grundlage für die Optimierung der Zuordnung individueller Präferenzfaktoren-Sets erscheint eine feinere Differenzierung zwischen

unterschiedlichen *Typen* von Pronomen – Possessiva, Reflexiva etc. – hilfreich. Diese Anforderung ergibt sich aus dem Layout des unterliegenden Rahmenalgorithmus, aus dessen Perspektive individuelle anaphorische Ausdrücke die zur Entscheidung, d.h. zur referentiellen Disambiguierung anstehenden Entitäten sind; die originäre, allein in den formalen Eigenschaften der inhaltlichen Ergebnisstruktur verankerte Evaluationsdisziplin spiegelt diese Sicht nicht wider. Wie anhand der Ergebnisse der in Abschnitt 4 zu diskutierenden Evaluation ferner ersichtlich werden wird, erweisen sich die originären, klassenbezogenen *Precision*- und *Recall*-Maße zudem als vergleichsweise schlechte Indikatoren für die Performanz in der Interpretation von *Pronominalanaphern*. Nicht zuletzt mit Blick auf die typischen Anforderungen von Anwendungen ergibt sich die Notwendigkeit einer präziseren Aufschlüsselung dieser Teilergebnisse.

3.2 Anwendungsperspektive: nichtpronominale Substitute

Wenn die primäre Leistung von Algorithmen zur referentiellen Interpretation unter Rekurs auf *Äquivalenzklassen* koreferenter Vorkommen gemessen wird, dann bleibt ein wichtiges Problem offen. Aus der Perspektive vieler Anwendungen besteht die eigentliche Aufgabe darin, *nichtpronominale Substitute* für pronominal-anaphorische Vorkommen zu ermitteln. Die Güte, mit der dies bewerkstelligbar ist, wird in der originären Evaluationsdisziplin nicht ermittelt. Die äquivalenzklassen-bezogene Evaluation würde in Beispiel (1) die Wahl des Pronomens “*er*” (Satz 1) als Antezedens der Pronominalanapher “*er*” (Satz 2) als korrekt bewerten. Der Performanzwert, der für das Problem der Ermittlung eines *beliebigen* kospezifizierenden, zur selben Äquivalenzklasse gehörigen Antezedens erzielt wird, kann jedoch nicht einfach auf das Problem der Ermittlung *nichtpronominaler* Anker für Pronomen übertragen werden. Vielmehr ergeben sich aus theoretischer Sicht Anhaltspunkte dafür, daß letztere Aufgabe schwieriger ist. Entsprechend den Vorhersagen einschlägiger Theorien der lokalen Fokussierung beziehen sich Pronominalanaphern tendenziell auf Diskursreferenten, die im jeweiligen Diskurszustand *fokussiert* sind, da in pronominalen Ausdrücken i.d.R. zu wenig Information enthalten ist, um ein Objekt außerhalb des Fokus eindeutig zu spezifizie-

ren. Unter der Prämisse kohärenten Diskurses ist ferner davon auszugehen, daß der Fokus des Diskurses nicht zu häufig wechselt. Folglich steht zu erwarten, daß die Entscheidung für ein pronominales, alle Restriktionen erfüllendes Antezedens mit höherer Wahrscheinlichkeit korrekt ist als die Entscheidung für einen nichtpronominalen Vorgänger. Hinzu kommt, daß die algorithmische Identifikation von Pronominalanapher und nichtpronominaler Okkurrenz als Mitglieder ein und derselben Koreferenzklasse in vielen Fällen auf einer (nichttrivialen) *Kette* individueller Antezedensentscheidungen basiert, wobei ein Fehler an beliebiger Stelle hinreichend für ein inkorrektes Ergebnis ist. In der Disziplin der Koreferenzklassen-Ermittlung ergäbe sich für folgende Wiederaufgriffs-Sequenz nur *ein* Precision-Fehler, jedoch wären sämtliche ermittelten nichtpronominalen Substitute inkorrekt.

(2) *Peter Behrens* \leftarrow^- *er* \leftarrow^+ *seine* \leftarrow^+ *er* \leftarrow^+ *ihm*

Als weitere, durch die Anforderungen potentieller Anwendungen motivierte Evaluationsdisziplin sollte daher das erwartet schwierigere Problem der *Ermittlung lexikalischer Substitute* betrachtet werden.

Die entscheidungsrelevanten Entitäten sind pronominale Okkurrenzen, für die geeignete lexikalische Substitute zu bestimmen sind. Im Sinne der Evaluation als geeignet, d.h. als korrekt angesehen werden sollen beliebige nichtpronominale Vorkommen, die entsprechend der Schlüssel-Spezifikation derselben Koreferenzklasse angehören. Zumindest aus theoretischer Sicht erscheint diese Vorgabe problematisch, denn im Falle der in der Sprachphilosophie ausgiebig diskutierten sog. *opaken* bzw. *intensionalen Kontexte* wird der Wahrheitswert nicht alleine durch die Referenz, sondern auch durch den konkreten sprachlichen Ausdruck, d.h. den *Sinn* (im Sinne von Frege) determiniert. Jedoch lassen sich eine Reihe von theoretischen Argumenten anführen, denen zufolge dieses Problem in praktischen Anwendungen keine Rolle spielen sollte (STUCKARDT 1999, S. 239ff). Auch aus Gründen der methodischen Vereinfachung soll daher am oben beschriebenen Evaluationskriterium festgehalten werden.

Somit besteht der Raum der systemgenerierten Lösungen aus einer Menge von Paaren (P,L) , wobei P ein Pronomen und L eine nichtpronominale Okkurrenz ist. Es gilt die Randbedingung, daß für jedes Pronomen

P maximal ein Paar (P,L) enthalten ist, denn es können auch Pronomen ohne Zuordnung existieren. Die Bewertung der Interpretationsleistung in der Disziplin der Ermittlung lexikalischer Substitute hat vor dem Hintergrund der durch den Schlüssel vorgegebenen Koreferenzrelation zu erfolgen, die genau solche Paare (P,L) lizenziert, in denen P und L ein und derselben Äquivalenzklasse angehören. Ein geeigneter Ausgangspunkt für die Definition adäquater *Precision*- und *Recall*-Maße ergibt sich auf der Grundlage einer Kategorisierung der systemgenerierten Paare entsprechend den Spezifikationen des Schlüssels (vgl. Tabelle 1). Bezugspunkt der Klassifikation sind Okkurrenzen P , die vom System als *pronominal* erkannt werden. Die Einordnung in die angegebenen Mengen geschieht in Abhängigkeit davon, ob P im Referenzkorpus als Vorkommen spezifiziert ist, ob eine Zuordnung (P,L) vorgenommen wurde sowie (ggf.) ob auch das vorgeschlagene Substitut L Vorkommenstatus im Schlüssel hat. Entsprechend den obigen Festlegungen wird eine Zuordnung (P,L) genau dann als *korrekt* angesehen, wenn P und L im Referenzkorpus als koreferent klassifiziert sind, wodurch impliziert wird, daß sowohl P als auch L im Schlüssel als Okkurrenzen ausgezeichnet sind. Im Hinblick auf die zu formulierenden Definitionen von *Precision*- und *Recall*-Maßen wird ferner zwischen *inkorrekten* und *leeren* Zuordnungen unterschieden, wobei jeweils feiner danach differenziert wird, ob die beteiligten Vorkommen Pendants im Schlüssel haben oder nicht.

Menge	Klassifikation	Definition
o_{++}	korrekt	P und L gehören im Schlüssel derselben Koreferenzklasse an
o_{+-}	inkorrekt	P und L gehören unterschiedlichen Schlüsselklassen an
$o_{+?}$	inkorrekt	P , nicht jedoch L entspricht einem Vorkommen im Schlüssel
o_{+_-}	leer	P entspricht Vorkommen im Schlüssel, kein Substitut L ermittelt
o_{+*}	leer	P entspricht Vorkommen im Schlüssel, kein Substitut L ermittelt, Koreferenz von P im Schlüssel als optional vorgegeben
$o_{?+}$	inkorrekt	P hat keine Entsprechung im Schlüssel
$o_{?-}$	leer	P hat keine Entsprechung im Schlüssel, kein Substitut L ermittelt

Tabelle 1: Klassifikation der ermittelten Pronomensubstitute

Anhand der Aufschlüsselung wird deutlich, daß auch hier Fehler in der Bestimmung objektspezifizierender Vorkommen die Qualität der Ergebnisse

des Prozesses der Bestimmung nichtpronominaler Substitute potentiell beeinträchtigen. Erneut stellt sich die Frage, ob und ggf. wie die beiden Evaluationsdisziplinen voneinander abgegrenzt werden sollten. Folgende Vorgehensweise erscheint adäquat: Die Betrachtungen werden eingeschränkt auf die eigentlichen entscheidungsrelevanten Entitäten, d.h. Pronomen P , die (auch) im Schlüssel als Okkurrenzen ausgezeichnet sind; relevant sind demnach ausschließlich die Mengen o_{++} , o_{+-} , $o_{+?}$, o_{+_-} und o_{+*} . Geeignete *Precision*- und *Recall*-Maße für die Aufgabe der Ermittlung nichtpronominaler Substitute ergeben sich durch folgende Festlegungen:

$$P_{ps} = \frac{|o_{++}|}{|o_{++}| + |o_{+-}| + |o_{+?}|}, \quad R_{ps} = \frac{|o_{++}|}{|o_{++}| + |o_{+-}| + |o_{+?}| + |o_{+_-}|}.$$

Da die Grundmenge der entscheidungsrelevanten Entitäten in Gestalt der systemseitig ermittelten pronominalen Okkurrenzen mit Schlüssel-Entsprechung fixiert ist und ausschließlich die Qualität der für die Instanzen dieser Menge zu treffenden Entscheidungen zur Evaluation ansteht, ergibt sich in bezug auf die angegebenen Definitionen die Besonderheit, daß stets $P_{ps} \geq R_{ps}$ ist. Die beschriebene Unterscheidung zwischen *Precision* und *Recall* erscheint aber dennoch angemessen, da ein *inneres* (auf die beschriebene Grundmenge begrenztes) Austauschverhältnis besteht, das durch die Beziehung zwischen der Menge o_{+_-} auf der einen und den Mengen o_{++} , o_{+-} und $o_{+?}$ auf der anderen Seite determiniert wird. Die definierten Maße sind somit aussagefähig im Hinblick auf die zentrale Frage, ob durch einen gezielten Verzicht auf bestimmte, als unsicher angesehene Entscheidungen eine Erhöhung der *Precision* erreicht werden kann.

4. Evaluation: Anaphernresolutions-System ROSANA

Das Anaphernresolutions-System ROSANA basiert auf einem Algorithmus, in dem *Restriktionen* zur Elimination zweifelsfrei inkorrekturer Kandidaten gefolgt von *Präferenzheuristiken* (Faktoren) zur Wahl unter den übrigen Kandidaten zum Einsatz gelangen. Während diese u.a. von Carbonell und Brown vorgeschlagene Rahmenstrategie in zahlreichen

Koreferenzalgorithmen Verwendung findet², strebt ROSANA zusätzlich die *uneingeschränkt robuste, massendatentaugliche Verarbeitung unrestringierter Texte* an.³ Es wird ausschließlich auf solche Ressourcen zur morphologischen und syntaktischen Voranalyse zurückgegriffen, die diese Voraussetzung ebenfalls erfüllen (STUCKARDT 1999, S. 243ff.). ROSANA setzt spezielle Techniken zur Arbeit auf *partiellen (fragmentarischen) syntaktischen Beschreibungen* ein, die insbesondere die repräsentationale Grundlage für eine zentrale Anaphernresolutions-Strategie - oberflächenstrukturelle konfigurationale Bedingungen - bilden (STUCKARDT 1997).

4.1 Diskussion der Evaluationsergebnisse

Abbildung 3 faßt die Ergebnisse der Evaluation des Anaphernresolutions-Systems ROSANA auf einem Korpus von 35 englischsprachigen Pressemeldungen zusammen (STUCKARDT 1999, S. 260ff.). Die Evaluation erfolgte unter Anwendungsbedingungen, d.h. ohne vorherige manuelle Korrektur orthographischer und syntaktischer Fehler in den Texten.

Der obere Teil von Abbildung 3 zeigt die Resultate in der Disziplin der Ermittlung referenzierender Vorkommen (vgl. Abschnitt 3.1.1); diese Aufgabe wird von ROSANA mit einer *Precision* von 0.9404 und einem *Recall* von 0.9623 gelöst. Die Ergebnisse indizieren eine bereits recht hohe Performanz. Eine qualitative Aufschlüsselung insbesondere der *Precision*-Fehler gibt näheren Aufschluß über das bestehende Verfeinerungspotential und legt offen, daß für die aus der Anwendungsperspektive besonders relevanten Pronominal-Okkurrenzen insbesondere im *Recall* überdurchschnittliche Ergebnisse erzielt werden (ibid., S. 262f.).

Im zweiten Teil von Abbildung 3 werden die Resultate in der originären, koreferenzklassen-bezogenen Evaluation nach dem MUC-Standard wiedergegeben; demnach liegt die *Precision* bei 0.8081 und der *Recall* bei 0.6845. Anhand der tabellarischen Aufschlüsselung nach unterschiedlichen Vorkommestypen wird die vergleichsweise geringe Aussagekraft des

² CARBONELL & BROWN (1988); vgl. etwa LAPPIN & LEASS (1994).

³ Entsprechend dieser Zielsetzung steht das Akronym ROSANA für "**R**obuste **s**yn-taxbasierte **I**nterpretation **a**naphorischer **A**usdrücke".

```

EVALUATIONSERGEBNIS OKKURRENZEN:
- NUR ANA: 243
- NUR KEY: 150
- ANA UND KEY: 3831
=> PRECISION: 0.9404
=> RECALL: 0.9623

A.R.-ERGEBNIS-PARTITIONIERUNG
- SCHNITTE: 256
- MOEGLICH: 1334
=> PRECISION: 0.8081

KEY-PARTITIONIERUNG
- SCHNITTE: 496
- MOEGLICH: 1572
=> RECALL: 0.6845

ANKNUEPFUNGS-ENTSCHEIDUNGEN

```

		PRECIS	++	+-	+?	+ ₋	+*	?+	? ₋
PRON	PE-3	0.7143	145	48	10	1	0	18	0
	PE12	0.9474	18	1	0	7	6	0	0
	PO-3	0.7634	100	28	3	0	0	0	0
	PO12	1.0000	3	0	0	1	1	0	0
	REFL	1.0000	3	0	0	1	0	0	0
	RELA	0.7789	74	18	3	6	0	7	4
			0.7555	343	95	16	16	7	25
NOMN	VNOM	0.7014	357	136	16	1973	43	31	133
	NAME	0.9390	308	15	5	368	5	5	28
			0.7945	665	151	21	2341	48	36

```

DURCHSCHNITT JE ANAPHER
=> PRECISION: 0.7808

LEXIKALISCHE SUBSTITUTION

```

		PRECIS	RECALL	++	+-	+?	+ ₋	+*	?+	? ₋
PE-3		0.6766	0.6667	136	54	11	3	0	18	0
PE12		0.9091	0.3846	10	1	0	15	6	0	0
PO-3		0.6641	0.6641	87	39	5	0	0	0	0
PO12		1.0000	0.5000	2	0	0	2	1	0	0
REFL		1.0000	0.7500	3	0	0	1	0	0	0
RELA		0.7667	0.6832	69	18	3	11	0	7	4

```

DURCHSCHNITT JE PRONOMEN
=> PRECISION: 0.7009
=> RECALL: 0.6532

```

Abbildung 3: Evaluationsergebnisse (Korpus von Pressemeldungen)

Precision-Werts in bezug auf die aus der Anwendungssicht wichtigere Interpretationsqualität für pronominale Ausdrücke deutlich: Im Durchschnitt liegt der Korrektheitsgrad auf der Ebene individueller Entscheidungen bei nur 75.55 %. Die Aufgliederung verdeutlicht ferner, daß etwa Reflexivpronomen (REFL; aufgrund des syntaktisch eingegrenzten Bezugskontexts) und textdeiktische Pronomen (PE12, PO12; aufgrund des distinktiven [kongruenzwirksamen] morphologischen Merkmals *Person*) mit wesentlich höherer Genauigkeit resolvierbar sind als herkömmliche Pronomen in dritter Person (PE-3, PO-3; nichtpossessiv bzw. possessiv).

Der untere Teil von Abbildung 3 zeigt die Ergebnisse in der Disziplin der Ermittlung nichtpronominaler Substitute. Entsprechend den Vorüberlegungen in Abschnitt 3.2 fällt die erzielte durchschnittliche *Precision* mit 0.7009 deutlich geringer aus als die *Precision* sowohl in der originären Disziplin (0.8081) als auch auf der Ebene der individuellen Antezedens-Entscheidungen (0.7555).

Die Betrachtungen belegen die Notwendigkeit und den großen Nutzen der Ergänzung der klassenbezogenen Evaluationsdisziplin sowie der verfeinerten Aufschlüsselung der Resultate.

4.2 Optimalität / Korpusabhängigkeit der Präferenzfaktoren

Die verfügbaren, mit referentieller Information ausgezeichneten Referenzkorpora wurden ferner dazu herangezogen, die anapherntypspezifischen Präferenzstrategien zur Antezedensauswahl zu optimieren sowie deren Korpusabhängigkeit zu untersuchen (vgl. Abschnitt 3.1.2). Die Basiskonfiguration der Faktormengen sowie der relativen numerischen Gewichtungen der Auswahlheuristiken von ROSANA wurde anhand eines *Trainingskorpus* von Pressemeldungen mit etwa gleicher Größe wie die der formalen Evaluation zugrundeliegenden Textsammlung optimiert.

Experiment	Evaluationskorpus "Pressemeldungen"						Korpus "Mozartopern"					
	P _{ÄK}	R _{ÄK}	P _{PS}	R _{PS}	PE3	PO3	P _{ÄK}	R _{ÄK}	P _{PS}	R _{PS}	PE3	PO3
ROSANA	0.81	0.68	0.70	0.65	0.71	0.76	0.88	0.81	0.75	0.74	0.79	0.77
(1) -SYR	0.80	0.68	0.68	0.63	0.70	0.73	0.89	0.82	0.76	0.75	0.77	0.80
(2) -SUP,...	0.80	0.68	0.67	0.62	0.69	0.73	0.87	0.80	0.68	0.68	0.74	0.67
(3) -SYR,-SUP,...	0.78	0.66	0.58	0.54	0.56	0.69	0.87	0.80	0.70	0.70	0.73	0.79
(4) -SDM	0.78	0.66	0.60	0.56	0.63	0.60	0.84	0.77	0.55	0.55	0.55	0.50
(5) -KAM	0.80	0.68	0.66	0.61	0.65	0.77	0.88	0.81	0.71	0.71	0.75	0.67

Tabelle 2: Variation der Präferenzstrategien von ROSANA

Tabelle 2 faßt die Ergebnisse von *fünf Experimenten mit den Präferenzfaktoren für Pronomen* auf zwei unterschiedlichen Evaluationskorpora (Pressemeldungen und Inhaltsbeschreibungen von Mozartopern) zusammen: (1) ohne syntaktische Rollenträgheit, (2) ohne Subjektpräferenz und ohne grammatische Rollenhierarchie, (3) = (1)+(2), (4) ohne Abwertung weiter entfernter Kandidaten (Abstandskriterium deaktiviert), (5) ohne Abwertung kataphorischer Wiederaufgriffe. In der obersten Zeile sind die Ergebnisse von ROSANA in der anhand des Trainingskorpus optimierten Originalkonfiguration wiedergegeben. Die Spalten PE-3 und PO-3 zeigen die spezifischen *Precision*-Werte der individuellen Antezedensentscheidungen für nichtpossessive bzw. possessive Pronomen in dritter Person; in den Spalten P_{PS} und R_{PS} sind ferner die durchschnittlichen *Precision*- und *Recall*-Ergebnisse in der Disziplin der Ermittlung lexikalischer Substitute angegeben.

Anhand der Werte in den Spalten P_{ÄK} und R_{ÄK} (koreferenzklassenbezogene *Precision*- und *Recall*-Werte) wird zunächst erneut die *unzureichende Sensitivität der originären Evaluationsmaße in bezug auf die Performanz für pronominale Anaphern* ersichtlich, denn im Vergleich zu den Größen P_{PS}, R_{PS}, PE-3 und PO-3 schlagen sich die Strategiemodifikationen in allenfalls marginaler Weise nieder. Dies ist ein weiteres deutliches Indiz für den Ergänzungsbedarf des MUC-Evaluationskanons gerade auch aus der Anwendungsperspektive, in der die am häufigsten auftretenden Dritte-Person-Pronomen im Zentrum des Interesses stehen.

Die Ergebnisse auf dem Evaluationskorpus "Pressemeldungen" verdeutlichen, daß die anhand eines Trainingskorpus desselben Textgenres

ermittelte Strategiekonfiguration tatsächlich (relativ zum eingeschränkten Skopus der vorgenommenen Variationen) zu optimalen Ergebnissen führt. Darüberhinaus ergeben sich weitere interessante Beobachtungen betreffend die Rollen der unterschiedlichen Faktoren. Dem algorithmisch trivialen Abstandskriterium kommt anscheinend die größte Bedeutung zu (Experiment (4)). Für Nichtpossessiva (PE-3) scheinen ferner die Faktoren "syntaktische Rollenträgheit" und "Subjektpräferenz / syntaktische Rollenhierarchie" in einem Substitutionsverhältnis zu stehen: Der Vergleich der Ergebnisse der Experimente (1) und (2) einerseits sowie (3) andererseits verdeutlicht, daß sich die Deaktivierung eines der beiden Faktoren in einer vergleichsweise moderaten Reduktion der Performanz niederschlägt, wohingegen die Ausblendung beider Faktoren zugleich zu erheblichen Einbußen führt.

Anhand einer Gegenüberstellung mit den Resultaten aus dem Korpus "Mozartopern" wird ferner ersichtlich, daß sich diese Ergebnisse nur in eingeschränktem Umfang generalisieren lassen. Einerseits wird die zentrale Rolle des Abstandskriteriums (Experiment (4)) bestätigt. Daß der Beitrag des Distanzfaktors sogar noch erheblich größer ausfällt, scheint seine Ursache in der spezifischen *Kohäsionsstruktur* der Operntexte zu finden: Der *lokale Fokus* wird durch die Protagonisten der jeweils beschriebenen Opernszene konstituiert; wechselt die Szene - was häufig geschieht -, so verschiebt sich der Fokus und damit der bevorzugte referentielle Anknüpfungspunkt. Unter Verzicht auf das Distanzkriterium führen die fokuspräferenz-approximierenden Heuristiken von Subjektpräferenz, syntaktischer Rollenträgheit und Rollenhierarchie somit häufiger zu einem falschen Ergebnis als für das Korpus "Pressemeldungen", in dem der lokale Fokus weniger häufig wechselt. Andererseits wird der positive Beitrag des Faktors "syntaktische Rollenträgheit" für Possessiva *nicht* bestätigt: Erneut findet sich die Erklärung mit Blick auf die charakteristische Kohärenzstruktur der Operntexte, in denen Belege mit multiplen Possessiva unterschiedlicher Referenz wie der folgende an der Tagesordnung sind:

- (3) *On a dark night in Seville, Leporello is keeping watch, grumbling, outside a house in which his master Don Giovanni is engaged in his latest amorous pursuit.*

Die Ergebnisse dieser Studie legen eine *korpuspezifische Optimierung der Auswahlstrategien* nahe. Auf der Grundlage von referentiell annotierten Korpora anwendungsrelevanter Textgenres wird es möglich sein, diesen Prozeß per computergestützter Analyse der spezifischen Kohärenzstrukturen zu automatisieren. Erneut wird das Potential korpusbasierter, formaler Evaluationen sowie der entsprechenden Ressourcen für Systementwicklung und Optimierung ersichtlich.

5. Zusammenfassung

Die Untersuchungen belegen die zentrale Bedeutung von Ressourcen zur formalen, korpusbasierten Evaluation für die Förderung der Entwicklung tatsächlich robuster, massendatentauglicher Textanalysetechnologie. Die Definition adäquater numerischer Maße für die Bewertung sowie den Vergleich der Interpretationsqualität von Softwaresystemen in relevanten Inhalterschließungsdisziplinen ist hierbei nur eine Seite der Medaille. Entscheidende Beiträge ergeben sich darüberhinaus für bestimmte Sekundärziele, die sich aus den Anforderungen von Systemdesign und Anwendung ableiten. Anhand des Problems der Evaluation von Systemen zur Koreferenzresolution wurden entsprechende Vorschläge zur Ergänzung der originären Evaluationsdisziplin erarbeitet. Eine erste Anwendung des erweiterten Disziplinenkanons im Rahmen der Evaluation des Anaphernresolutions-Systems ROSANA hat interessante Anhaltspunkte für mögliche Richtungen einer erfolgversprechenden Optimierung der Interpretationsstrategien ergeben.

Anhand der Ausführungen wird deutlich, daß die formale, korpusbasierte Evaluation von Softwaresystemen zur Tiefenanalyse von Texten ein komplexes Metier ist, dem der Status einer eigenständigen Forschungsdisziplin zukommt. Die zentrale Relevanz dieses Themas wurde erneut belegt. Folglich sollten die notwendigen Schritte initiiert werden, diese Disziplin sowie das Textanalyseparadigma *Information Extraction* im deutschsprachigen Raum besser zu etablieren. Insbesondere könnte dies die Schaffung eines entsprechenden *Pools von Ressourcen* (annotierte deutschsprachige Korpora, sprachspezifische Aufgabendefinitionen, Annotations- und Evaluationssoftware, robuste Komponententechnologie) sowie

eines *institutionalen Rahmens* (Evaluationskonferenzen) beinhalten. Eine derartige Zentralisierung sollte sich als wichtiger Motor der Weiterentwicklung deutschsprachiger Texttechnologie bewähren.

Literaturverzeichnis

- CARBONELL, Jaime G. and BROWN, Ralf D. (1988): Anaphora Resolution: A Multi-Strategy Approach. In: *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest 1988, 96-101.
- CoTD 1997: MUC-7 Coreference Task Definition, Version 3.0, 13.7.97. Bezogen über: Nancy Chinchor, SAIC (chinchor@gso.saic.com). Auch in: *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998, veröffentlicht unter <http://www.muc.saic.com/>.
- GRISHMAN, Ralph and SUNDHEIM, Beth (1996): Design of the MUC-6 Evaluation. In: *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, San Francisco 1996, 1-11. Morgan Kaufmann.
- HIRSCHMAN, Lynette (1998): The Evolution of Evaluation. Lessons from the Message Understanding Conferences. In: *Computer Speech and Language* 12, 281-305.
- LAPPIN, Shalom and LEASS, Herbert J. (1994): An Algorithm for Pronominal Anaphora Resolution. In: *Computational Linguistics* 20 (4), 535-561.
- STUCKARDT, Roland (1997): Resolving Anaphoric References on Deficient Syntactic Descriptions. In: *Proceedings of the ACL'97 / EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, Madrid, July 1997, 30-37.
- (1999): Qualitative Inhaltsanalyse durch Computer – ein uneinlösbarer Anspruch? – Untersuchungen zur algorithmischen Textinhaltserschließung am Beispiel der referentiellen Interpretation. Dissertation, Fachbereich Gesellschaftswissenschaften, Johann Wolfgang Goethe-Universität Frankfurt am Main, Januar 1999.
- SUNDHEIM, Beth (1996): Overview of Results of the MUC-6 Evaluation. In: *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, San Francisco 1996, 13-31. Morgan Kaufmann.
- VILAIN et al. 1996: Marc V., John BURGER, John ABERDEEN, Dennis CONNOLLY, Lynette HIRSCHMAN, A Model-Theoretic Coreference Scoring Scheme. In: *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, San Francisco 1996, 45-52. Morgan Kaufmann.