Robuste Anaphernresolution

Roland Stuckardt

1 Einleitung

Ein zentrales Teilproblem der computergestützten Erschließung und strukturierten Aufbereitung des Inhalts digitalisiert vorliegender Textdokumente besteht in der referenziellen Interpretation anaphorischer Ausdrücke. Folglich sind robuste, operationale Verfahren zur Anaphernresolution, die unter Anwendungsbedingungen auf einem breiten Spektrum von Texten arbeiten, von hoher softwaretechnologischer Relevanz. Entsprechend groß ist die Aufmerksamkeit, die diesem Thema seit Mitte der 90er Jahre in der sprachtechnologischen Forschung entgegengebracht wird. Mit Blick auf die einschlägige Literatur wird deutlich, dass die Entwicklung robuster Inhaltserschließungssoftware von der Verfügbarkeit standardisierter texttechnologischer Ressourcen entscheidend profitiert.[1]

Im vorliegenden Beitrag sollen die wesentlichen Probleme der robusten, operationalen Anaphernresolution identifiziert und eine Übersicht des Stands von Forschung und Technologie gegeben werden. Besonderes Augenmerk soll der zentralen Rolle einer standardisierten Texttechnologie für Entwicklung, Optimierung und Evaluation robuster Softwarelösungen gelten. Die Ergebnisse der Studie sind allgemeiner Art insofern, als sie für ein breites Spektrum von Textinhaltserschließungsproblemen gelten.

2 Zum Problem: Anaphern- bzw. Koreferenzresolution

2.1 Warum Anaphern- bzw. Koreferenzresolution essenziell ist

Am Beispiel der Inhaltserschließungsdisziplin *Information Extraction* (IE) lässt sich die zentrale Bedeutung der referenziellen Interpretation anaphorischer Ausdrücke illustrieren. Das Ziel der IE besteht darin, inhaltliche Entitäten bestimmter Klassen, die in einer vorgegebenen Anwendungsdomäne relevant sind, in unstrukturiertem Text algorithmisch zu erkennen und strukturiert aufzuberei-

1 In Bezug auf die Anaphernresolution vgl. u.a. den Tagungsband (Mitkov und Boguraev 1997) eines ACL-Workshops, das Themenheft (Mitkov, Boguraev und Lappin 2001) der Zeitschrift *Computational Linguistics* sowie die Monographie (Mitkov 2002).

ten.[2] Im Allgemeinen geht es um Entitäten komplexen Zuschnitts wie z.B. Fakten, durch welche Entitäten mit Objektbezug in Relation gesetzt werden. Zum Teil spiegeln sich diese Entitäten unmittelbar in der Prädikat-Argument-Struktur einzelner Teilsätze wieder und sind deshalb auf der Grundlage einer syntaktischen Analyse erschließbar. Im Allgemeinen jedoch verteilen sich die sprachlichen Ausdrücke, welche die Mitspieler sowie die sie verbindenden Aussagen realisieren, auf mehrere u.U. weit auseinander liegende Sätze; sie zu erschließen und repräsentational zu integrieren, bedarf deshalb der Anwendung von Algorithmen zur Analyse des textglobalen Diskurses.

Anhand eines typischen Beispiels aus der fünften Message Understanding Conference (1995) soll dies illustriert werden. Die Zielvorgabe für die zu entwickelnden IE-Softwaresysteme bestand darin, aus textuellem Input des Typs Wirtschaftsmeldung faktuelle Information über Joint-Venture-Vereinbarungen zu extrahieren. Anhand von folgendem zielinhaltlich relevantem Textausschnitt ist ersichtlich, dass die korrekte Interpretation bestimmter anaphorischer Ausdrücke für die Erschließung der Joint-Venture-Fakten essenziell ist:[3]

> *Bridgestone Sports Co. said Friday **it** has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan. **The joint venture**, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and metal wood clubs a month.*

Satz 1 kommuniziert einen Joint-Venture-Sachverhalt. Dieser spiegelt sich explizit in der syntaktischen Prädikat-Argument-Struktur eines Komplementsatzes wider. Jedoch ist es erst vermöge der referenziellen Identifikation des Pronomens ***it*** mit dem Antezedensvorkommen *Bridgestone Sports Co.* möglich, eine den Zielvorgaben der IE-Aufgabenstellung genügende *Nichtpronominalform* für den pronominal realisierten Partizipanten des Joint-Venture-Sachverhalts zu substituieren. Mit Blick auf den zweiten hervorgehobenen Ausdruck wird ferner deutlich, dass auch die Bezüge nichtpronominaler Anaphern potenziell relevant sind: Auf der Grundlage der referenziellen Identifikation der definiten NP ***The joint venture*** (Satz 2) und *a joint venture* (Satz 1) wird erkennbar, dass Satz 2 zusätzliche Informationen zum Joint-Venture-Faktum beisteuert, dessen Beschreibung in Satz 1 begonnen wurde.

---

2 Somit lässt sich Informationsextraktion generisch beschreiben als die Aufgabe, bestimmte Textinhalte zu erkennen und derart aufzubereiten, dass sich diese in einer entsprechend strukturierten relationalen Datenbank ablegen lassen.

3 Hervorhebungen der näher betrachteten anaphorischen Ausdrücke durch den Autor

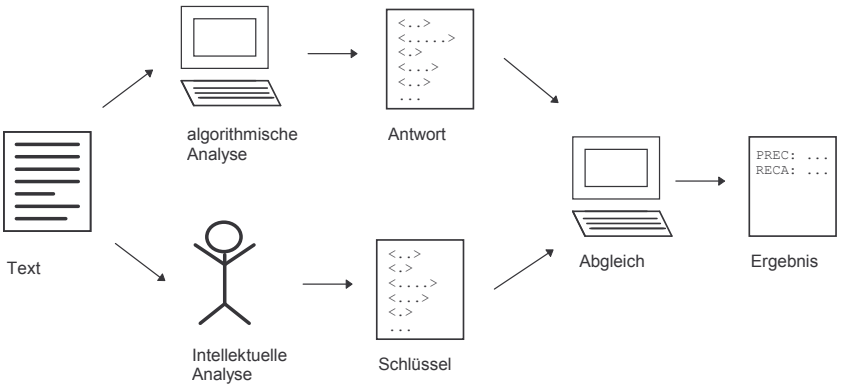

Abbildung 1 Szenario der formalen, korpusbasierten Evaluation

2.2 Koreferenzresolution als eigenständige Inhaltserschließungsdisziplin

Die grundlegende Bedeutung, die somit der Interpretation anaphorischer Ausdrücke im Rahmen jedweder weiter gehenden Erschließung des textuellen Inhalts zukommt, wurde zum Anlass genommen, das Problem der Koreferenzresolution als eigenständige Inhaltserschließungsdisziplin zu betrachten. Im Rahmen der Message Understanding Conferences 6 und 7 wurden entsprechende inhaltliche Zielvorgaben für Softwarelösungen zur Koreferenzresolution erarbeitet.[4] Die formale Evaluation der Systeme erfolgt per Abgleich der systemseitig identifizierten Klassen koreferenter sprachlicher Ausdrücke mit Referenzklassen, die durch menschliche Annotatoren erzeugt werden. Die Interpretationsleistung wird gemäß einem Schema bewertet, das den Grad der Übereinstimmung der systemgenerierten und der intellektuell erzeugten Klassen misst, wobei nach Precision und Recall unterschieden wird (Vilain et al. 1996). Ausgehend von der Beobachtung, dass dieses Schema in Bezug auf die zentrale Klasse *pronominaler* Anaphern wenig aussagefähige Ergebnisse liefert, wurden weitere Evaluationsmaße definiert, welche die Interpretationsleistung in den Teildisziplinen der Identifikation unmittelbarer Antezedenten und Konzeptsubstitute für Pronomen bewerten (Stuckardt 2001). Weiter gehende Vorschläge zur Standardisierung der Evaluation von Systemen zur (pronominalen) Anaphernresolution werden von Byron (2001) unterbreitet.

2.3 Korpusbasierte Evaluation mithilfe standardisierter Texttechnologie

4 Vgl. die Konferenzproceedings (MUC-6 1996) und (MUC-7 1998) sowie (Hirschmann 1998).

Es soll nun zunächst eine Übersicht der im Rahmen der Evaluation von Koreferenz- und Anaphernresolutionssystemen zum Einsatz gelangenden *standardisierten Texttechnologie* gegeben werden. Unabhängig von der jeweiligen Inhaltserschließungsdisziplin erfolgt die Evaluation i.d.R. gemäß dem in Abbildung 1 skizzierten generischen Schema. Demnach wird das Ergebnis der computergestützten Analyse in einem standardisierten Repräsentationsformat *(Antwort)* abgelegt; selbiges gilt für die intellektuell erzeugten Referenzdaten *(Schlüssel)*. Die Evaluation der Systemperformanz erfolgt per automatischem, softwaregestütztem *Abgleich* der systemgenerierten Ergebnisse mit den Vorgaben des Ergebnisschlüssels. Im Rahmen der Koreferenz-Evaluation der Message Understanding Conferences 6 und 7 galten einheitliche Formatstandards für die Repräsentation von Systemergebnissen und Referenzdaten. Die referenzielle Information wird direkt im Originaltext annotiert. Für den oben betrachteten Text sähe eine inhaltlich richtige referenzielle Auszeichnung folgendermaßen aus:

> *<COREF ID="1">Bridgestone Sports Co.</COREF> said <COREF ID="2">Friday</COREF>* ***<COREF ID="3" TYPE="IDENT" REF="1">it</COREF>*** *has set up <COREF ID="4">a joint venture</COREF> in <COREF ID="5">Taiwan</COREF> with <COREF ID="6">a local concern</COREF> and <COREF ID="7">a Japanese trading house</COREF> to produce <COREF ID="8">golf clubs</COREF> to be shipped to <COREF ID="9">Japan</COREF>.* ***<COREF ID="10" TYPE="IDENT" REF="4">The joint venture</COREF>****, <COREF ID="11" TYPE="IDENT" REF="10">Bridgestone Sports Taiwan Co. </COREF>, capitalized at <COREF ID="12">20 million new Taiwan dollars</COREF>, will start <COREF ID="13">production</COREF> in <COREF ID="14">January 1990</COREF> with <COREF ID="15" TYPE="IDENT" REF="13">production</COREF> of <COREF ID="16" TYPE="IDENT" REF="8">20,000 iron and metal wood clubs</COREF> a <COREF ID="17">month</COREF>.*

Die zu annotierenden Basisentitäten sind objektreferenzierende sprachliche Ausdrücke (sog. *Vorkommen/Okkurrenzen*); gemäß Annotationsschema sind diese in die Marken *<COREF ID=...>* und *</COREF>* einzuschließen. Jedem Vorkommen wird in der öffnenden Marke eine textweit eindeutige Identifikationsnummer zugeordnet (*ID="..."*). Beziehungen referentieller Identität werden per Angabe der Identifikationsnummer eines beliebigen koreferenten Antezedensvorkommens *(REF="...")* vermerkt. Ferner ist der Typ der Referenz einzutragen; da bislang ausschließlich der mehr oder weniger elementare[5] Fall der referenziellen *Identität* (*Ko*referenz) betrachtet wird, steht stets *TYPE="IDENT"*. Auf der Grundlage der in diesem Format beschriebenen Ketten anaphorischer Wiederaufgriffe ergeben sich per Berechnung des

5 Die Erfahrungen aus den Message Understanding Conferences 6 und 7 belegen, dass die Objektkoreferenzinterpretation einige schwierige Probleme aufwirft. Die versteckte Komplexität schlägt sich in einer überraschend niedrigen Interannotator-Übereinstimmung von 84% nieder, die im Rahmen der intellektuellen Annotation der Referenzkorpora gemessen wurde.

transitiven Abschlusses auf elementare Weise die Koreferenzklassen. Somit kann das klassenbezogene Evaluationsschema von Vilain et al. (1996) angewendet werden, um Systemergebnis und Schlüsseldaten abzugleichen.

Der erste, grundlegende Beitrag einer standardisierten Texttechnologie besteht also darin, die Evaluation robuster Software-Basistechnologie für die Inhaltserschließung nach formalen Kriterien zu ermöglichen:

1. Auf der Grundlage der inhaltlichen Eingrenzung der Inhaltserschließungsdisziplin sowie der formalen Definitionen von Evaluationsmaßen (i.d.R. Precision, Recall) lässt sich die Performanz von Softwaresystemen *individuell* auf methodisch abgesicherte, reproduzierbare Weise bewerten.
2. Die Standardisierung und zentrale Verfügbarmachung einer entsprechenden Texttechnologie, welche *Aufgabenbeschreibungen, Annotationsschemata und zugehörige Werkzeuge, inhaltlich annotierte Textkorpora, Definitionen formaler Evaluationsmaße sowie zugehörige Bewertungssoftware* umfasst, ermöglicht eine *vergleichende* Evaluation unterschiedlicher Softwarelösungen für dieselbe Inhaltserschließungsaufgabe, die aussagefähige, reproduzierbare Resultate liefert.

Vermöge der formalen, korpusbasierten Evaluation werden die methodischen Mängel überwunden, die einer isolierten intellektuelle Evaluation von Algorithmen unter möglicherweise idealisierenden Rahmenbedingungen anhaften. Es wird gewährleistet, dass die Verfahren *uneingeschränkt operational* sind und *robust* auf anwendungsrelevanten, nicht voredierten Texten arbeiten, dass ferner *Skopus* sowie *Aussagekraft* der Performanzmaße expliziert sind und dass die Evaluationsergebnisse unterschiedlicher Systeme *vergleichbar* sind.

Eine Reihe weiterer Vorteile ergeben sich in Bezug auf die *Entwicklung* von Softwaresystemen zur Textinhaltserschließung. Um eine Basis für die Diskussion dieses Punkts am Beispiel der Anaphernresolution zu schaffen, soll zunächst ein genauerer Blick auf die Interpretationsstrategien geworfen werden, die typischerweise Verwendung finden.

3 Operationale Strategien für die Anaphernresolution

Für die algorithmische Interpretation anaphorischer Ausdrücke sind eine Vielzahl von Strategien vorgeschlagen worden. Grundbedingung ist die *Operationalität*: Es ist sicher zu stellen, dass die Strategien auf beliebigen Texten der jeweiligen Anwendungsdomäne laufen und dass die Wissensquellen, auf denen in den Strategien rekurriert wird, in dem benötigten Umfang zur Verfügung stehen. Relational aufgeschlüsselt schlägt sich die Operationalitätsbedingung in

der Forderung von *Robustheit gegenüber* eine Reihe typischer Problemfälle nieder:[6]

- Robustheit gegenüber orthographischen und grammatikalischen Fehlern, d.h. der Inhaltserschließungsalgorithmus sollte auch sprachlich in Teilen fehlerhafte Texte ‚möglichst gut' verarbeiten können;
- Robustheit gegenüber partieller Verfügbarkeit strategierelevanten Wissens – wird etwa auf syntaktische Beschreibungen rekurriert, so sollte der Algorithmus auf fragmentarischen (partiellen oder ambigen) Syntaxbäumen arbeiten, da vollständige und eindeutige Syntaxanalysen unter nichtidealisierenden Verarbeitungsbedingungen eher die Ausnahme denn die Regel sind.

Neben Operationalität ist ferner zu fordern, dass die jeweilige Strategie einen *hinreichend generellen Skopus* aufweist. Sie sollte einen wesentlichen Beitrag für die Koreferenzinterpretation anwendungsrelevanter Texte leisten und nicht nur in wenigen, möglicherweise trivialen Spezialfällen greifen.

Gemeinhin wird davon ausgegangen, dass die *Identifikation* der objektreferenzierenden sprachlichen Ausdrücke (d.h. das Setzen der Okkurrenzmarken *<COREF ID="..."> ... </COREF>)* bereits vollzogen ist und entsprechende Okkurrenzrepräsentationen angelegt sind. Im Folgenden werden somit eine Reihe von Strategien diskutiert, die einen Beitrag zur Anaphernresolution im engeren Sinne (d.h. dem Einfügen inhaltlich richtiger Koreferenzattribute (… *REF="..."* …) in die Marken anaphorischer Ausdrücke) leisten.

Carbonell und Brown (1988) schlagen eine grundlegende Unterscheidung von Strategien gemäß ihrer jeweiligen Stringenz vor: Sog. *Restriktionen* geben Evidenz, welche Antezedenskandidaten für eine bestimmte anaphorische Okkurrenze *definitiv nicht* in Frage kommen; sog. *Präferenzen* machen heuristische Aussagen darüber, welche der die Restriktionen erfüllenden Kandidatenvorkommen für eine bestimmte Anapher *mit hoher Wahrscheinlichkeit korrekt* sind. Beispiele robust operationalisierbarer restriktiver Strategien mit hinreichendem Grad an Allgemeingültigkeit sind *Kongruenz* in Person, Numerus und (teilweise) Genus sowie *syntaktisch-konfigurationale Zulässigkeit*:

(1) *Der Architekt betrat das Büro seiner Kollegin.*
***Er** diskutierte mit **ihr** die überarbeiteten Pläne.*
(2a) *Der Kunde verlangt, dass der Friseur **sich** rasiert.*
(2b) *Der Kunde verlangt, dass der Friseur **ihn** rasiert.*

6 Vgl. Menzel (1995), der in Bezug auf die Robustheit von Sprachverarbeitungssystemen drei unabhängige Kriterien identifiziert: *Monotonie* (je weniger defizient der Input, desto besser die Interpretationsleistung), *Autonomie* (einzelne Analysemodule arbeiten möglichst autark; ein Scheitern der Verarbeitung auf einer Ebene sollte nicht das Scheitern der Verarbeitung auf weiteren Ebenen bedingen) und *Interaktion* (Analysemodule unterschiedlicher Ebenen sollten einander zuarbeiten, um etwaige lokale Defizienzen global zu kompensieren).

(2c) *Der Kunde verlangt, dass der Friseur* ***den Kunden*** *rasiert.*

Beleg *(1)* illustriert die Wirksamkeit der Kongruenzbedingung: Während das Pronomen ***Er*** nur mit dem Ausdruck *Der Architekt* koreferieren kann, kann das Pronomen ***ihr*** ausschließlich auf die durch den Ausdruck *seiner Kollegin* in den Diskurs eingeführten Entität Bezug nehmen; der inhaltliche Hintergrund liegt in der Reflexion des natürlichen Geschlechts der Diskursreferenten im grammatischen Geschlecht der an der sprachlichen Oberfläche realisierten Ausdrucksformen. Die robuste Operationalisierbarkeit ist gewährleistet, da lexikalische Ressourcen mit Genusinformation sowie darüber hinaus robuste Algorithmen zur morphologischen Analyse mit hohem Abdeckungsgrad zur Verfügung stehen.[7] Aus empirischen Experimenten ist ferner bekannt, dass diese Strategie einen wesentlichen Beitrag für die algorithmische Koreferenzresolution leistet.

Die Belege *(2a)*, *(2b)* und *(2c)* illustrieren die Wirksamkeit syntaktisch-konfigurationaler Bedingungen, die bestimmte satzinterne Distributionsmuster referenzidentischer sprachlicher Ausdrücke ausschließen. Für das Reflexivpronomen ***sich*** in Beleg *(2a)* etwa besteht die Bedingung der Koreferenz mit einer Entität, die innerhalb des Nebensatzes sprachlich realisiert wird – somit ist Koreferenz mit dem ‚lokalen' Antezedens *der Friseur* obligatorisch; betreffend das nichtreflexive Pronomen ***ihn*** in Beleg *(2b)* hingegen sind (in erster Näherung) nebensatzlokale Antezedenten ausgeschlossen; in Bezug auf die nichtpronominale Anapher ***den Kunden*** in Beleg *(2c)* erscheinen darüber hinaus bestimmte satzinterne Antezedenskandidaten außerhalb des Nebensatzes konfigurational unzulässig. Phänomene dieser Art werden beispielsweise in den Bindungsprinzipien der Government and Binding (GB) Theory formalisiert, die auf konfigurationalen Eigenschaften der syntaktischen Oberflächenstruktur aufsetzen (vgl. Chomsky 1981). Folglich setzt die Operationalisierung dieser Restriktionsklasse voraus, dass syntaktische Beschreibungen verfügbar sind. Da syntaktische Information i.d.R. nur partiell vorliegt, sind Maßnahmen zu ergreifen, um diese Restriktionen auf fragmentarischen syntaktischen Beschreibungen oder auf noch wissensärmeren Part-of-Speech-Auszeichnungen robust zu operationalisieren.[8] Lösungen dieses Problems werden von Kennedy und Boguraev (1996) und von Stuckardt (2001, 2000, 1997) erarbeitet. Dass die syntaktisch-konfigurationalen Kriterien einen wesentlichen Beitrag für die algorithmische Koreferenzresolution leisten, ist aus empirischen Studien bekannt (Lappin und Leass 1994).

---

7 Zusätzliche Arbeit verbleibt in Bezug auf Eigennamen zu leisten, denn ein System zur morphologischen Analyse verfügt i.d.R. nicht über die Genusinformation zu Namensausdrücken.

8 Robuste Syntaxanalysatoren (Järvinen und Tapanainen 1997) bzw. Part-Of-Speech-Tagger (Karlsson et al. 1995) sind für eine Reihe von Sprachen verfügbar.

Den Präferenzstrategien kommt die Rolle zu, aus den nicht vermöge stringenter operationaler Kriterien ausschließbaren Kandidatenokkurrenzen eine Auswahl zu treffen. Erneut ist auf Operationalisierbarkeit und Abdeckungsgrad zu achten; aus ersterem Grund kann daher nicht auf solche Präferenzstrategien zurückgegriffen werden, die auf tiefeninhaltlichen Repräsentationen bzw. Inferenzen aufbauen. Übliche Kriterien setzen daher auf oberflächenstrukturell-syntaktischen Beschreibungen auf oder rekurrieren auf statistische Daten aus der Analyse großer (annotierter oder nichtannotierter) Korpora. Sie sind deshalb Ersatzverfahren für inhaltsadäquate Entscheidungsstrategien und insofern heuristischer Natur.[9] Zu den wichtigsten Präferenzstrategien, die in robusten Anaphernresolutionsverfahren (i.d.R. in kombinierter Form) zum Einsatz kommen, gehören die *Subjektpräferenz* sowie das *Rollenträgheitskriterium*:[10]

(3a) *Der Kunde besucht den Friseur.* ***Er*** *parkt vor dem Salon.*
(3b) *Der Architekt besucht den Kollegen.* ***Er*** *trifft* ***ihn*** *im Büro.*

In Beleg *(3a)* etwa könnte die Heuristik der Subjektpräferenz (mangels Verfügbarkeit tiefeninhaltlicher Kriterien) den Ausschlag dafür geben, dem Personalpronomen ***Er*** das in der syntaktisch prominenten Subjektfunktion realisierte Antezedens *Der Kunde* zuzuordnen. In Beleg *(3b)* könnte es das Kriterium der Rollenträgheit nahe legen, für die beiden Pronominalanaphern ***Er*** und ***ihn*** jeweils diejenigen Antezedenten (*Der Architekt* bzw. *den Kollegen*) zu wählen, die dieselben syntaktischen Rollen (Subjekt bzw. transitives Objekt) bekleiden; eine Wahl des Subjektantezedenten *Der Architekt* für beide Pronomen würde in diesem Fall übrigens alleine schon deshalb nicht zum Ziel führen können, weil die Koreferenz dieser beiden Anaphern aufgrund der oben diskutierten syntaktisch-konfigurationalen Bedingungen ausgeschlossen ist. Weitere Präferenzfaktoren, die in operationalen Systemen zum Einsatz gelangen, sind *Hierarchie der grammatischen Funktion* (Subjekt vor direktem Objekt vor indirektem Objekt vor PP-Objekt) sowie die in der Literatur ausgiebig diskutierten *Centering-Strategien*, die auf einer elaborierten Theorie der lokalen Kohärenz basieren.[11]

9 Ersichtlich wird dies etwa anhand der von Dagan und Itai (1990) vorgeschlagenen lexikalisch-semantischen Kookkurrenz-Präferenzstrategie, die auf statistischen Korpusdaten aufsetzt.

10 Die hier bzw. im Folgenden skizzierten Präferenzstrategien für Personalpronomen sind insofern fokus- bzw. kohäsionstheoretisch motiviert, als Kandidaten in syntaktisch hervorgehobenen Rollen bevorzugt bzw. – allgemeiner – solche Muster referenziellen Wiederaufgriffs präferiert werden, die den gegenwärtigen (lokalen) referenziellen Fokus beibehalten. Die empirische Rechtfertigung solcher Strategien wird u.A. in psycholinguistischen Experimenten untersucht.

11 Vgl. u.a. Grosz, Joshi und Weinstein 1995. Betreffend die Centering-Strategien steht der Beweis der Vorteilhaftigkeit gegenüber elementareren präferenzfaktorenbasierten Modellen aus. Zudem sind die Centering-Regeln in ihrer originären Form unterspezifiziert; bislang scheint es nicht gelungen, die bestehenden Freiheitsgrade in empirisch zufrieden stellender Form zu schließen.

Somit ergibt sich folgendes *Rahmenschema* für Verfahren zur Anaphern- bzw. Koreferenzresolution:

1. Für alle Anaphern A: Identifikation derjenigen Antezedenskandidaten K(A), die alle Restriktionen relativ zu A erfüllen.
2. Für alle Anaphern A: Ermittlung einer Plausibilitätsordnung der verbliebenen Kandidaten K(A) vermöge der heuristischen Präferenzstrategien.
3. Für alle Anaphern A: Wahl des jeweils plausibelsten Kandidaten $K^*(A)$, der allen unmittelbaren und mittelbaren Bedingungen genügt.

4 Manuelle Optimierung robuster Verfahren zur Anaphernresolution

Eine ganze Reihe von Algorithmen zur Anaphernresolution basiert auf dem im vorangegangenen Abschnitt vorgestellten Rahmenschema, in dem zwischen möglichst früh anzuwendenden Restriktionen und im Vorfeld der Auswahlentscheidung zum Einsatz gelangenden Präferenzfaktoren unterschieden wird.[12] Ein gemeinsames Merkmal dieser Ansätze besteht darin, dass die Auswahl und Optimierung der Interpretationsstrategien *intellektuell* erfolgt. In Bezug auf den Mix an Präferenzfaktoren erweist sich dieser Prozess als aufwändig und demnach wert einer Unterstützung durch geeignete texttechnologische Werkzeuge.

Somit ergibt sich ein weiteres zentrales Einsatzgebiet für inhaltlich (hier: referenziell) annotierte Korpora, Evaluationsstandards und Bewertungssoftware. Steht eine standardisierte Texttechnologie zur Verfügung, so wird der Prozess der schrittweisen Verfeinerung des Strategiemixes erheblich vereinfacht. Beginnend mit der Wahl eines geeigneten, dem geplanten Anwendungsgebiet möglichst nahen Trainingskorpus und einer Grundkonfiguration an Strategien wird die Performanz des Softwaresystems in Bezug auf die Grundkonfiguration formal und *automatisch* evaluiert. Aufbauend auf einer intellektuell-qualitativen Analyse der Fehlerfälle wird sodann der Strategiemix zielgerichtet verfeinert und der Evaluationszyklus beginnt von vorne. Inhaltlich annotierte Korpora und die entsprechenden Software-Evaluationswerkzeuge entlasten somit den Systementwickler in der Reevaluierungsphase wesentlich.

Im Rahmen der inhaltlichen Eingrenzung und Standardisierung von Evaluationsdisziplinen ist deshalb nicht nur darauf zu achten, dass den in Abschnitt 2.3 identifizierten Anforderungen der individuellen bzw. vergleichenden Bewertung der Inhaltserschließungsperformanz Rechnung getragen wird. Von ähnlich großer Bedeutung ist die Aussagefähigkeit der Evaluationsmaße in Bezug auf bestimmte Eckdaten der Systemperformanz, die per Variation einzelner Strategieparameter bzw. Strategien beeinflusst werden können. Beispielsweise hat es sich gezeigt, dass die für die MUC-Evaluationen entwickelten koreferenzklas-

12 Vgl. die auf vollständigen syntaktischen Beschreibungen aufsetzenden idealisierenden Verfahren von Hobbs (1978) und Lappin und Leass (1994) sowie die robusten Algorithmen von Kennedy und Boguraev (1996), Baldwin (1997) und Stuckardt (2001, 2000) (ROSANA-System).

senbezogenen Evaluationsmaße zur Bewertung der Performanz in der Interpretation von Pronomen bzw. spezifischer Typen anaphorischer Ausdrücke nicht hinreichend sensitiv sind (Stuckardt 2001). Um als Werkzeug für die Feinabstimmung der oft anapherntypspezifischen Restriktionen und Präferenzfaktoren zu taugen, sollte das Evaluationsschema in geeigneter Weise untergliedert sein, d.h. Precision- und Recallwerte differenziert nach Vorkommenstypen ausweisen.

In Bezug auf das oben beschriebene Rahmenschema für Anaphernresolutionsalgorithmen hat sich gezeigt, dass die *Präferenz*strategien einer sorgfältigen Optimierung bedürfen. Idealerweise sollte der Präferenzfaktorenmix auf die Charakteristika des anwendungsspezifischen Text*genres* zugeschnitten werden: Empirische Studien haben Anhaltspunkte dafür ergeben, dass relatives Gewicht und Anwendungsskopus der Faktoren Subjektpräferenz und Rollenträgheit mit Blick auf die spezifische Kohäsionsstruktur des jeweiligen Genres festgelegt werden sollten (Stuckardt 2001). Einmal mehr zeigt sich der Vorteil der Verfügbarkeit einer Texttechnologie bestehend aus einer zentral vorgehaltenen Sammlung annotierter Korpora *unterschiedlicher Genres* sowie der zugehörigen Softwarewerkzeuge für eine möglichst feinkörnige Evaluation: Es ergeben sich neue, tiefgehende Einsichten betreffend den verbleibenden Spielraum zur Optimierung robuster Anaphernresolutionsalgorithmen.

5 Maschinelles Lernen robuster Interpretationsstrategien

Auch wenn durch den Einsatz inhaltlich annotierter Korpora eine wesentliche Unterstützung erzielt werden kann, so geht die systematische manuelle Optimierung von Anaphernresolutionsalgorithmen mit einem erheblichen intellektuellen Aufwand bei der qualitativen Analyse der Fehlerfälle und der sich anschließenden zielgerichteten Modifikation der Interpretationsstrategien einher. Gerade dann, wenn auf die aufwandsarme Entwicklung eines robusten, operationalen Verfahrens abgezielt wird – es also *nicht* darum geht, mit einer spezifischen Einzelstrategie unter ggf. idealisierten Rahmenbedingungen qualitativ-empirisch zu experimentieren –, steht ein alternatives Entwicklungsmodell zur Verfügung, in dem *Algorithmen des Maschinellen Lernens*[13] zum Einsatz kommen. Der Grundgedanke einiger zentraler Formen des Maschinellen Lernens (ML) besteht darin, auf der Basis einer repräsentativen Menge vorklassifizierter Trainingsdaten einen Entscheidungsalgorithmus zu lernen, der die ‚neuen' Daten des Anwendungsfalls mit hoher Genauigkeit klassifiziert.

Derartige ML-Verfahren können zum automatischen Lernen von Klassifikatorfunktionen für die Koreferenzresolution genutzt werden. Die entscheidungsrelevanten Datensätze bestehen aus Beschreibungen (Attributvektoren) v(A,K),

13 Vgl. z.B. (Mitchell 1997).

die über je zwei Okkurrenzen berechnet werden: einer zu resolvierenden Anapher A und einem Antezedenskandidaten K. Für dieses Paar (A,K) ist nun mit Blick auf dessen Repräsentation v(A,K) zu entscheiden, ob Koreferenz gegeben ist. Mit anderen Worten: Jedes Paar (A,K) ist in genau eine der beiden Klassen *KOREF* und *NICHT-KOREF* einzuordnen. Somit ist im Rahmen des Maschinellen Lernens eine Klassifikatorfunktion zu berechnen, die beliebigen Beschreibungen v(A,K) eine eindeutige Klassenkennzeichnung (KOREF oder NICHT-KOREF) zuordnet. Geeignete Trainingsdaten lassen sich auf der Grundlage annotierter Korpora berechnen: In einem Leerdurchlauf des jeweiligen Algorithmus werden Beschreibungen v(A,K) zu allen Anapher-Kandidat-Paaren (A,K) angelegt, über deren Koreferenz dieser Algorithmus im Anwendungsfalle zu befinden hätte. Per Lookup des referenziell annotierten Korpus wird ermittelt, welcher Klasse C der Vektor v(A,K) zugeordnet werden sollte; der entsprechend *klassifizierte* Attributvektor v(A,K)*:C* bildet einen Trainingsfall. Über dieser Menge von Trainingsdaten berechnet nun das jeweilige ML-Verfahren eine Klassifikatorfunktion, die im Anwendungsfall durch den Anaphernresolutionsalgorithmus konsultiert wird, um für die Beschreibung v(A',K') eines beliebigen Paares (A',K') heuristisch zu entscheiden, ob Koreferenz gegeben ist.

Die Signatur[14] der Vektoren v(A,K) ist dahingehend festzulegen, dass die Instanzen der zugrunde liegenden Attributmengen robust berechenbar sind. Bezüglich der zur Verfügung stehenden Information gelten somit dieselben Einschränkungen wie für das manuelle Systemdesign; jedoch fällt es in den Aufgabenbereich des ML-Verfahrens, die im Rahmen des zu erlernenden Klassifikationsalgorithmus tatsächlich relevante Teilmenge von Attributen *automatisch* zu ermitteln. Die Güte des erlernten Klassifikators hängt von der zur Verfügung stehenden Attributinformation ab – in Abwesenheit umfassender semantischer Information ist folglich nicht damit zu rechnen, dass die Interpretationsqualität erheblich über derjenigen manuell konfigurierter Resolutionsalgorithmen liegt. Jedoch liegt ein potenzieller Vorteil des Maschinellen Lernens darin, dass verborgene Regularitäten aufgespürt werden, die einem möglicherweise theoretisch voreingenommenen menschlichen Systemdesigner entgehen.

In Bezug auf die Koreferenzresolution sieht demnach die ML-Strategie folgendermaßen aus: Wähle eine *möglichst umfassende* Menge von Vorkommensattributen, die über der verfügbaren oberflächenpositionalen, lexikalischen/morphologischen, syntaktischen und semantischen Information *robust* berechenbar sind; definiere hierüber die Signatur der zu generierenden Attributvektoren; unter der Bedingung, dass eine hinreichend große, repräsentative Menge von Trainingsfällen zur Verfügung steht, ermittelt nun der Lernalgorith-

14 Unter Signatur sei das Kartesische Produkt $A_1 \times A_2 \times \ldots \times A_n$ der Attributmengen verstanden.

mus, welche der zur Verfügung stehenden Attribute mit welcher relativen bzw. bedingten Wichtigkeit in den Entscheidungsprozess einfließen sollten.

Das zuvor beschriebene ML-Modell für die Anaphernresolution liegt einer Reihe von Verfahren zugrunde, die in den letzten Jahren entwickelt wurden. Zu nennen sind u.a. die Arbeiten von Conolly, Burger und Day (1994), Aone und Bennett (1995), Soon, Ng und Lim (2001) sowie Stuckardt (2002). Der Ansatz von Aone und Bennett (1995) etwa zielt auf die Interpretation anaphorischer Ausdrücke des Japanischen ab. Als ML-Verfahren wird der Algorithmus *C4.5* von Quinlan (1993) eingesetzt, mit Hilfe dessen auf der Grundlage des oben beschriebenen Rahmenverfahrens Klassifikatoren in Form von Entscheidungsbäumen oder (mittelbar) Mengen von Entscheidungsregeln gelernt werden. Aone und Bennet verwenden C4.5, um einen Entscheidungsbaum für die Resolution von vier Klassen anaphorischer Formen des Japanischen (Namen, definite NPs, ‚*zero pronouns*', ‚*quasi-zero pronouns*') über Vektoren mit einer Signatur bestehend aus 66 lexikalischen, syntaktischen, semantischen und oberflächenpositionalen Attributen zu lernen. Der erlernte Klassifikator verkörpert eine Generalstrategie, der an die Stelle sowohl der Restriktionen als auch der Präferenzfaktoren klassischer, manuell konfigurierter Systeme tritt. Soon, Ng und Lim (2001) verfolgen einen ähnlichen Ansatz für englischsprachige Texte. Als Lernverfahren findet die C4.5-Weiterentwicklung C5 Verwendung; die betrachteten Eigenschaftsvektoren umfassen lediglich 12 robust berechenbare Attribute. Per Evaluation auf annotierten Korpora von MUC-6 und MUC-7, zu denen die Ergebnisse manuell konfigurierter Systeme bekannt sind, gelingt der Nachweis, dass ML-basierte Anaphernresolutionssysteme eine Interpretationsqualität erzielen können, die derjenigen der handkonfigurierten Algorithmen ebenbürtig ist.

Indes besteht ein wesentlicher Unterschied zwischen den beiden Grundklassen von Strategien, der eine fokussierte Anwendung der ML-Technik im Rahmen der Anaphernresolution nahe legt. Eine strategiebezogene Evaluation hat ergeben, dass die oben beschriebenen stringenten Strategien der Kongruenz sowie der syntaktisch-konfigurationalen Bedingungen nahezu verlustfrei robust operationalisierbar sind (Stuckardt 2001). Des Weiteren wurde beobachtet, dass es die Präferenzstrategien sind, die in Abhängigkeit vom anwendungsspezifischen Textgenre konfiguriert werden sollten (vgl. Abschnitt 4). Wenn jedoch die *Restriktionen* mit genre- und domänenübergreifender Gültigkeit operationalisiert werden können, dann bietet es sich an, maschinell gelernte Klassifikationsfunktionen als genrespezifisch festzulegendes Substitut des *Präferenz*strategienmixes und *in Ergänzung* der global gültigen und deshalb nur einmal manuell zu implementierenden stringenten Strategien einzusetzen. Dieses Vorgehensmodell liegt dem Ansatz (Stuckardt 2002) zugrunde, der ebenfalls auf dem ML-Algorithmus C4.5 basiert und auf Eigenschaftsvektoren mit bis zu 38 Attributen (lexikalische/morphologische, syntaktische und oberflächenpositionale Informa-

tion) arbeitet. In einer empirischen Evaluation dieses Verfahrens (ROSANA-ML) auf einem englischsprachigen Korpus wurde eine Interpretationsqualität nachgewiesen, die das Niveau des manuell optimierten Vorgängersystems ROSANA erreicht. Da die Trainingsmengen relativ klein waren, wird Spielraum für weitere Steigerungen gesehen.

In Abbildung 2 ist das Zusammenspiel von Lernphase und Anwendungsphase ML-basierter Anaphernresolutionssysteme am Beispiel von ROSANA-ML skizziert. Die in der Lernphase mit Hilfe des jeweiligen ML-Algorithmus gelernte Klassifikationsfunktion wird in der Anwendungsphase konsultiert, um Koreferenzentscheidungen zu treffen. Anhand der Abbildung wird die zentrale Rolle einer standardisierten Texttechnologie auch im Rahmen des maschinellen Lernens von Anaphernresolutionsstrategien deutlich. Annotierte Korpora fließen einerseits in den Trainingsprozess und andererseits in die Anwendungs- und Evaluationsphase ein; die Bewertung erfolgt unter Rückgriff auf die Evaluationssoftware. Die Verfügbarkeit annotierter Korpora unterschiedlicher Textgenres erleichtert es, die Präferenzstrategien von Anaphernresolutionsalgorithmen auf den spezifischen Anwendungsfall hin abzustimmen und entsprechende Gemeinsamkeiten bzw. Unterschiede der für die Genres charakteristischen Kohäsionsstrukturen empirisch zu untersuchen.

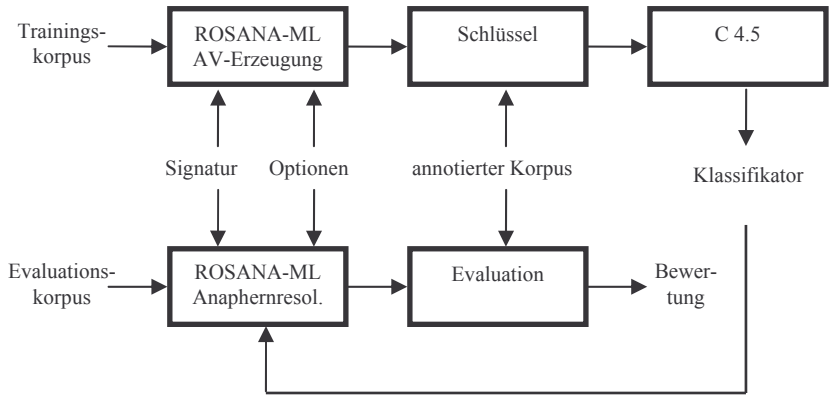

Abbildung 2 ROSANA-ML: Lernphase und Anwendungsphase

6 Zusammenfassung und Ausblick

Anaphernresolution ist ein zentrales Teilproblem der computergestützten Erschließung des Inhalts digitalisiert vorliegender Textdokumente. Um zu ge-

währleisten, dass entsprechende Algorithmen unter den Bedingungen des Anwendungsfalls operational sind, ist auf Interpretationsstrategien zurückzugreifen, die gegenüber defizientem Input – orthographisch/grammatisch nicht fehlerfreien Texten, partiellen syntaktischen/semantischen Repräsentationen – *robust* sind. Im vorliegenden Beitrag ist am Beispiel der Anaphernresolution gezeigt worden, dass die Entwicklung operationaler Inhaltserschließungsverfahren zumindest in viererlei Hinsicht von einer standardisierten Texttechnologie, welche Aufgabendefinitionen, Annotationsschemata und zugehörige Werkzeuge, annotierte Korpora, Definitionen formaler Evaluationsmaße sowie entsprechende Bewertungssoftware umfasst, profitiert. Die standardisierte Texttechnologie

1. ermöglicht die methodisch abgesicherte, reproduzierbare Bewertung der individuellen Performanz nachweislich lauffähiger, operationaler Softwaresysteme unter Anwendungsbedingungen;
2. setzt Maßstäbe für den validen, reproduzierbaren Vergleich unterschiedlicher Softwaresysteme unter identischen Bedingungen;
3. unterstützt den Systementwickler bei der manuellen Verfeinerung von Interpretationsstrategien sowie (mittelbar) bei der empirischen Exploration inhaltlicher Regularitäten von Textgenres bzw. Anwendungsdomänen;
4. erlaubt die Anwendung von Algorithmen des Maschinellen Lernens zum automatischen Ableiten robuster Interpretationsstrategien, da annotierte Korpora zur Generierung von Trainingsdaten genutzt werden können.

Somit hängen Fortschritte in der Entwicklung operationaler Verfahren zur Textinhaltserschließung in weiten Zügen vom Ausbau zentral vorgehaltener texttechnologischer Ressourcen ab. Folglich sollten Anstrengungen unternommen werden, weitere Inhaltserschließungsdisziplinen zu definieren und entsprechende Ressourcen für eine größere Bandbreite von Sprachen verfügbar zu machen.

In Bezug auf die Anaphernresolution dürfte der Schwerpunkt zukünftiger Forschungen auf den Gebieten Maschinelles Lernen bzw. Statistik, d.h. im *automatischen* Design operationaler Interpretationsstrategien liegen. Die Möglichkeiten der manuellen Systemkonfiguration scheinen ausgereizt zu sein.[15] Es verbleibt zu sehen, inwieweit sich die empirisch motivierten Erwartungen erfüllen, die an die automatische genrespezifische Konfiguration der Interpretationsstrategien mit Verfahren des Maschinellen Lernens geknüpft werden.

15 Um einmal konkrete Zahlen zu nennen: Das State-of-the-Art-System ROSANA (Stuckardt 2001, 2000) erreicht auf Texten des Genres ‚Pressemeldungen' bezüglich Pronominalanaphern 76% Korrektheit (Precision) in der Wahl eines beliebigen (auch pronominalen) Antezedens und 70% Korrektheit in der Wahl eines nichtpronominalen lexikalischen Substituts.

7 Literatur

Aone, Chinatsu/ Bennett, Scott William (1995). Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies. In: Proceedings of the 33rd Annual Meeting of the ACL, Santa Cruz, New Mexico. 122-129.

Baldwin, Breck (1997). CogNIAC: High Precision Coreference with Limited Knowledge and Linguistic Resources. In: Mitkov, Ruslan/ Boguraev, Branimir (Hrsg.) (1997): 38-45.

Byron, Donna (2001). The Uncommon Denominator: A Proposal For Consistent Reporting of Pronoun Resolution Results. In: Computational Linguistics 27(4). 2001. 569-578.

Carbonell, Jaime G./ Brown, Ralf D. (1988). Anaphora Resolution: A Multi-Strategy Approach. In: Proceedings of the 12th International Conference on Computational Linguistics (COLING). 1988. 96-101.

Chomsky, Noam (1981). Lectures on Government and Binding. Dordrecht: Foris Publications.

Conolly, Dennis/ Burger, John D./ Day, David S. (1994). A Machine-Learning Approach to Anaphoric Reference. In: Proceedings of the International Conference on New Methods in Natural Language Processing (NEMLAP).

Dagan, Ido/ Itai, Alon (1990). Automatic Processing of Large Corpora for the Resolution of Anaphoric References. In: Proceedings of the 13th International Conference on Computational Linguistics (COLING). 1990. 330-332.

Grosz, Barbara J./ Joshi, Aravind K./ Weinstein, Scott (1995). Centering: A Framework for Modeling the Local Coherence of Discourse. In: Computational Linguistics 21(2). 1995. 203-225.

Hirschmann, Lynette (1998). MUC-7 Coreference Task Definition, Version 3.0. In: MUC-7 (1998): ohne Seitenangabe.

Hobbs, Jerry R. (1978). Resolving Pronoun References. In: Lingua, 44, 1978: 311-338.

Järvinen, Timo/ Tapanainen, Pasi (1997). A Dependency Parser for English. Technical Report TR-1. Helsinki: University of Helsinki, Department of General Linguistics.

Karlsson, Fred/ Voutilainen, Atro/ Heikkila, Juha/ Antilla, Arto (1995). Constraint Grammar: A Language-Independent System For Parsing Free Text. Berlin/ New York: Mouton de Gruyter.

Kennedy, Christopher/ Boguraev, Branimir (1996). Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING). 1996. 113-118.

Lappin, Shalom/ Leass, Herbert J. (1994). An Algorithm for Pronominal Anaphora Resolution. In: Computational Linguistics 20(4). 1994. 535-561.

Menzel, Wolfgang (1995). Robust Processing of Natural Language. In: Wachsmut, Ipke/ Rollinger, Claus-Rainer/ Brauer, Wilfried (Hrsg.) (1995). KI-95: Advances in Artificial Intelligence: 6th Annual German Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence 981, Berlin/ Heidelberg/ New York: Springer Verlag. 19-34.

Mitchell, Tom M. (1997). Machine Learning. New York: McGraw-Hill.

Mitkov, Ruslan (2002). Anaphora Resolution. Oxford: Longman.

Mitkov, Ruslan/ Boguraev, Branimir (Hrsg.) (1997). Proceedings of the ACL'97 / EACL'97 Workshop on Operational Factors in Practical, Robust Anaphor Resolution for Unrestricted Texts, Madrid. Somerset, NJ: Association for Computational Linguistics (ACL).

Mitkov, Ruslan/ Boguraev, Branimir/ Lappin, Shalom (Hrsg.) (2001). Computational Lingistics: Special Issue on Computational Anaphora Resolution. Vol. 27(4). 2001.

MUC-6 (1996). Proceedings of the 6th Message Understanding Conference (MUC-6). San Francisco: Morgan Kaufmann.

MUC-7 (1998). Proceedings of the 7th Message Understanding Conference (MUC-7). Unveröffentlicht, ehemals (9.12.1999) verfügbar unter http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html.

Quinlan, J.R. (1993). C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann.

Soon, Wee Meng/ Ng, Hwee Tou/ Lim, Daniel Chung Yong (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. In: Computational Linguistics 27(4). 2001. 521-544.

Stuckardt, Roland (1997). Resolving Anaphoric References on Deficient Syntactic Descriptions. In: Mitkov, Ruslan/ Boguraev, Branimir (Hrsg.) (1997): 30-37.

Stuckardt, Roland (2000). Qualitative Inhaltsanalyse durch Computer – ein uneinlösbarer Anspruch? Untersuchungen zur algorithmischen Textinhaltserschließung am Beispiel der referentiellen Interpretation. Berlin: Tenea-Verlag.

Stuckardt, Roland (2001). Design and Enhanced Evaluation of a Robust Anaphor Resolution Algorithm. In: Computational Linguistics 27(4). 2001. 479-506.

Stuckardt, Roland (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 2002). Lissabon: Edições Colibri. 211-216.

Vilain, Marc/ Burger, John/ Aberdeen, John/ Conolly, Dennis/ Hirschmann, Lynette (1996). A Model-Theoretic Coreference Scoring Scheme. In: MUC-6 (1996): 45-52.