DIE SUCHE NACH INFORMATIONEN UNTER SPRACHWISSENSCHAFTLICHEN GESICHTSPUNKTEN: DAS POTENZIAL VON ANAPHERN

von Helene Schmolz

Inhalt

Zusammenfassung: *Die Informatik tendiert dazu, zur Lösung von natürlichsprachigen Anwendungen vor allem quantitative Analysen zu verwenden sowie qualitative Methoden, die einfach umzusetzen sind. Doch könnte eine Einbeziehung der Sprachwissenschaft auf Augenhöhe solche Anwendungen entscheidend verbessern. Ein vielversprechender Ansatz sind Anaphern, die den Text semantisch-syntaktisch verknüpfen und dadurch den Inhalt eines Textes an der Oberfläche spiegeln. Auf Basis einer umfassenden Definition und Kategorisierung zeigt sich an einem Korpus, dass die Anaphernart, die bisher noch nie beachtet wurde, am häufigsten auftritt. Diese Erkenntnisse bergen großes Potenzial, wenn sie bei Suchmaschinen, wo Anaphern bisher generell kaum Beachtung finden, angewandt werden.*

Schlüsselwörter: *Text Retrieval, Anapher, Anaphernauflösung, Korpuslinguistik, Suchmaschinen, Ranking*

SEARCHING FOR INFORMATION FROM A LINGUISTIC POINT OF VIEW: THE POTENTIAL OF ANAPHORS

Abstract: *Information science tends to use predominantly quantitative analyses and also qualitative methods that are easy to implement, in order to solve natural language tasks. However, the integration of linguistics on an equal footing could improve such tasks decisively. One promising approach are anaphors, which connect the texts in a semantic-syntactic way and thereby reflect the content of texts on the surface. Based on a comprehensive definition and categorisation, it is shown by means of a corpus that the type of anaphor which has so far never been regarded is in fact the most frequent. This knowledge offers high potential if it is applied to search engines, where anaphors have generally attracted little attention to date.*

Keywords: *Text Retrieval, anaphor, anaphora resolution, corpus linguistics, search engines, ranking*

1. Einleitung

Suchdienste wie Google sind angesichts der heutigen Datenflut im Web nicht mehr wegzudenken. Die Funktionsweise von Suchmaschinen, im Besonderen, wie der Benutzer mit einer Suchanfrage möglichst relevante Dokumente zurückerhält, basiert auf einem mathematischen Verständnis. Dabei werden etwa Texte nicht nach anspruchsvollen inhaltlichen Kriterien durchsucht, sondern es wird mit einem einfachen Zählen von Wörtern auf die Relevanz dieses Textes geschlossen. Sucht ein Benutzer beispielsweise nach *Kaffee*, so wird ein Dokument, in dem *Kaffee* sechs mal auftritt, als relevanter erachtet, als eines, in dem es nur drei mal gezählt wurde, auch wenn sich bei näherer Betrachtung beider Texte genau das umgekehrte Bild ergeben würde.

Dass ein Suchdienst mit dieser einfachen Strategie mitunter viele Dokumente zurückgibt, die nicht oder nicht exakt zum gewünschten Suchbegriff passen, erleben wir tagtäglich. Daher scheint es zunehmend drängender, Disziplinen wie die Sprachwissenschaft mit einzubinden, um mit deren Methoden etwa den Inhalt von Texten besser repräsentieren zu können. Konkret wird hier ein Ansatz vorgestellt, wie die Berücksichtigung von sogenannten Anaphern helfen kann, Texte in englischer Sprache semantisch detaillierter analysieren zu können. Dazu wird zuerst erläutert, was Anaphern sind, welche Arten von Anaphern es im Englischen gibt und wie häufig diese Arten jeweils sind. Anschließend wird skizziert, wie Anaphernresolution funktioniert und welche Ansätze es bisher gibt. Schließlich wird erläutert, wie Suchmaschinen Texte analysieren und welches Potenzial eine Anaphernresolution bei Suchmaschinen hat.

2. Von Anaphern und was dahinter steckt

Anaphern werden in der Sprachwissenschaft Ausdrücke genannt, die auf einen meist vorher genannten Ausdruck, das sogenannte Antezedens, zurückverweisen. In Beispiel (1) ist die Anapher *er* enthalten, die auf ein Element im vorausgehenden Satz verweist, also *Thomas*. Dabei werden nur solche Elemente als Anaphern bezeichnet, bei denen das Antezedens im Text selbst verankert ist. Würde beispielsweise in (1) der erste Satz fehlen,

aber wäre aus der Situation erkennbar, um welche Person es sich handelt, würde *er* nicht als Anapher zählen. Die Zuordnung des richtigen Antezedens zu jeder Anapher wird schließlich als Anaphernauflösung oder Anaphernresolution bezeichnet.

(1) **Thomas** liest gerne. Er bevorzugt Krimis.

Dabei ist selbst in der Sprachwissenschaft umstritten, welche Ausdrücke nun konkret als Anaphern bezeichnet werden und welche nicht mehr dazu gehören. Noch seltener wurde bislang beachtet, wie Anaphern definiert werden sollten, damit die Definition auch in einem computergestützten System anwendbar ist. Nur auf Basis einer präzisen Definition kann anschließend auch eine umfassende Kategorisierung vorgenommen werden. Aus einer Definition unter sprachwissenschaftlichen und informationstechnologischen Aspekten (vgl. dazu Schmolz, Döller & Coquil 2012; Schmolz & Coquil 2014) ergeben sich 12 Arten von Anaphern für die englische Sprache: *central pronouns* (Beispiel 2); *reciprocal pronouns* (3); *demonstrative pronouns* (4); *relative pronouns* (5); *adverbs* (6); *noun phrases with a definite article* (7); *proper names* (8); *indefinite pronouns* (9); *other forms of coreference and substitution:* the same*,* such *and* so (10); *verb phrases with* do *and combinations with* so*,* this*,* that*,* it *and* the same (thing) (11); *ellipses* (12); *non-finite clauses* (13). Zur besseren Lesbarkeit wird jede Anapher unterstrichen und jedes Antezedens in den Beispielen fett gedruckt.

(2) **Susan** plays the piano. She likes music.
(3) **The children** told each other a story.
(4) Many people play **the guitar**. This instrument is probably the most popular one.
(5) Ann called out to **her friend Tom**, who was just crossing the street.
(6) Ms Smith was in **London**. She came back from there yesterday.
(7) He went by **car**. After a while, the engine broke down.
(8) **Bob Harris** is at a meeting in Berlin today. In urgent cases you can call the secretary there – just ask for Mr Harris.
(9) I need **a pen**. Do you have one?
(10) **Tina will come to the party.** At least I hope so.
(11) Mary **speaks English perfectly**. At least, I think she does.
(12) If you really have to buy a **guitar**, do not get the cheapest ___.
(13) **The apparatus** examining the heartbeat of new-borns attracts the attention of the experts.

Um festzustellen, wie häufig diese 12 Anaphernarten in Texten auftreten, wurde dies an einem selbst erstellten Hypertextkorpus untersucht. Das Korpus umfasst dabei Texte aus Wikipedia und Blogs, Webseiten von Tageszeitungen, Unternehmens-Webseiten, persönliche und institutionelle Webseiten und enthält insgesamt fast 76.000 Wörter. Die prozentuale Verteilung der Anaphernarten ist in Abbildung 1 dargestellt.

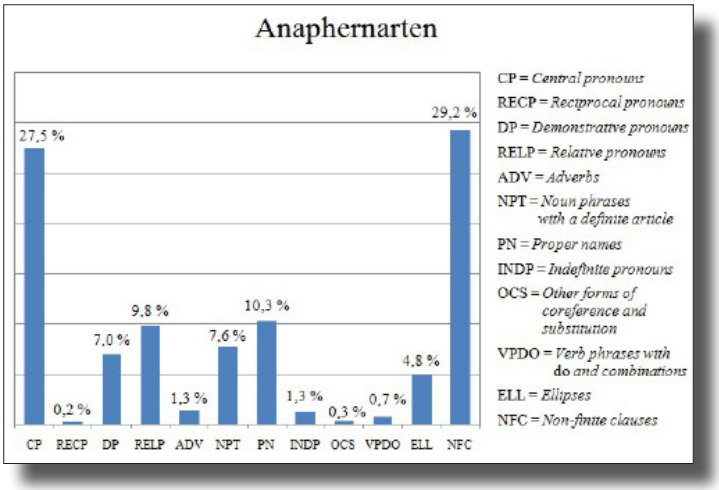

Abb. 1: Verteilung der Anaphernarten im Korpus

Die Analyse zeigt, dass die häufigste Anaphernart die *non-finite clause* Anaphern sind, was insofern überrascht, als dass diese bislang kaum beachtet und bis dato noch nie als eine Anaphernart klassifiziert wurden. Als häufigste und wichtigste galten vielmehr *central pronouns*, zu denen etwa Personalpronomen wie *she* oder *they* gehören. Insgesamt nehmen *central pronouns* und *non-finite clause* Anaphern mehr als die Hälfte aller Anaphern ein. Weiters von Bedeutung sind noch *proper names*, *relative pronouns*, *noun phrases with a definite article*, *demonstrative pronouns* und *ellipses*.

3. Funktionsweise einer Anaphernresolution und bisherige Ansätze

Anaphernresolutions-Systeme führen grundsätzlich drei Schritte durch. Zuerst wird nach Anaphernindizien gesucht und entschieden, ob tatsächlich eine Anapher vorliegt oder nicht (siehe Abbildung 2). Beispielsweise ist *she* ein Anaphernindiz, da es nicht bei jedem Auftreten im Text eine Anapher darstellt (vergleiche die Diskussion zu Beispiel 1). Im zweiten Schritt

erfolgt die Auswahl möglicher Antezedens-Kandidaten, bevor im dritten Schritt das korrekte Antezedens auf Basis verschiedener Regeln eruiert wird (vgl. Mitkov 2002: 18–19, 33–47). Im Beispiel kommen *Susan* und *the piano* als Antezedens in Frage. Da jedoch *she* meist auf eine feminine Person verweist, ist das Antezedens *Susan* wahrscheinlicher.

Schritte	**Anaphernresolution von** ***Susan plays the piano. She likes music.***
1. Anapherndetektion	*she* (anaphorisch)
2. Detektion von Antezedens-Kandidaten	*Susan, the piano*
3. Auswahl des wahrscheinlichsten Antezedens	*Susan* (da mit *she* in Genus übereinstimmend; beide sind feminin)

Abb. 2: Beispiel einer Anaphernresolution

Bisherige Ansätze zur Anaphernresolution können – in Analogie zu anderen Unterscheidungen, wie etwa bei Systemen der natürlichen Sprachverarbeitung – in zwei grundlegende Kategorien eingeteilt werden: regelbasiert und datenbasiert. Anfangs wurden vor allem regelbasierte Ansätze entwickelt. Seit den 1990er Jahren sind es vor allem datenbasierte Ansätze, die jedoch noch nicht so gute Ergebnisse erzielen wie regelbasierte (vgl. Mitkov 2002: 95; Mitkov & Hallett 2007: 271). Regelbasierte Ansätze sind arbeitsintensiver, da sie im Vorfeld mehr Wissen, das in Form von Regeln repräsentiert wird, benötigen (vgl. Strube 2010: 400–407). Einer der ersten regelbasierten, jedoch noch immer zitierten, ist Hobbs Algorithmus (vgl. Hobbs 1976). Außerdem sind bei regelbasierten Ansätzen Lappin & Leass (1994) und Haghighi & Klein (2009) zu nennen. Datenbasierte Ansätze benötigen hingegen ein Korpus, aus dem sie selbstständig Regeln ableiten. Wichtige datenbasierte Ansätze umfassen allem voran Soon, Ng & Lim (2001); daneben sind Versley et al. (2008), Stoyanov et al. (2010) und Uryupina (2010) erwähnenswert.

Alle existierenden Ansätze, ob regel- oder datenbasiert, beschränken sich jeweils auf bestimmte Anaphernarten; kein Ansatz umfasst alle Arten. Außerdem berücksichtigt kein Ansatz die so häufigen *non-finite clause* Anaphern. Erschwerend kommt hinzu, dass in den veröffentlichten Studien

oft nicht genau dargelegt wird, welche Anaphern oder Anaphernarten ein System eigentlich beabsichtigt, aufzulösen. Umso wichtiger ist es, verschiedene Anaphernresolutions-Systeme zu vergleichen. Zum Beispiel können Systeme mittels Precision und Recall evaluiert werden. Dabei wird unterschieden, ob nur die Anapherndetektion, also wie viele Anaphern korrekt erkannt werden, oder auch die anschließende Anaphernresolution, das heißt die Zuordnung des richtigen Antezedens zu jeder Anapher, bemessen werden. Für beide Fälle ist die Berechnung in Abbildung 3 und 4 angegeben (vgl. Baldwin 1997: 41–42).

$$\text{Precision} = \frac{\text{Gefundene Anaphern}}{\text{Insgesamt gefundene Anaphernindizien}}$$

$$\text{Recall} = \frac{\text{Gefundene Anaphern}}{\text{Anaphern insgesamt}}$$

Abb. 3: Anapherndetektion

$$\text{Precision} = \frac{\text{Anzahl der korrekt aufgelösten Anaphern}}{\text{Anzahl der Anaphern, die aufzulösen versucht wurde}}$$

$$\text{Recall} = \frac{\text{Anzahl der korrekt aufgelösten Anaphern}}{\text{Anzahl aller Anaphern}}$$

Abb. 4: Anaphernresolution

Um verschiedene Systeme vergleichen zu können, reichen diese Maße jedoch nicht aus. Wünschenswert ist es, unterschiedliche Systeme auf ein und demselben Korpus und mit gleichen Vorverarbeitungsschritten zur Aufbereitung des Textes zu testen. Dies zeigt sich insofern als Desiderat, da bisherige Systeme nicht nur verschiedene Korpora und Vorverarbeitungsschritte verwenden, sondern auch gewisse Vorverarbeitungsschritte beispielsweise nur simulieren oder anschließend Fehler manuell korrigieren, wie etwa Lappin & Leass (1994) (vgl. Mitkov & Hallett 2007: 262). Ein Benchmarking-System, das verschiedene Systeme mit gleichen Ressourcen testet, stammt von Mitkov & Hallett (2007: 262–263). Sie zeigen anhand von *central pronouns*, dass die von ihnen erzielten Werte viel niedriger sind als in den Studien angegeben wurde. Um eine solche umfassende Evaluation bei allen Anaphernarten durchführen zu können, ist ein frei zugängliches Korpus notwendig, das übrigens auch zur Entwicklung datenbasierter Ansätze verwendet werden könnte. Bisherige Korpora beschränken sich

– ähnlich wie Ansätze zur Anaphernresolution – auf bestimmte Anaphernarten und ignorieren *non-finite clause* Anaphern. Daher wurden alle Anaphern im selbst erstellten Hypertextkorpus annotiert. Dieses Korpus soll schließlich öffentlich zugänglich gemacht werden, sodass die zukünftige Forschung daran anknüpfen kann.

4. Wie Suchmaschinen Texte analysieren

Da hier allein die Suche nach Texten relevant ist, soll sich die Funktionsweise von Suchmaschinen auf eine Suche nach Texten beschränken. Allgemeiner gesprochen liegen dabei nun Text Retrieval Systeme vor, die aus einer Datenbank jeweils diejenigen Textdokumente heraussuchen, die der Benutzer gerade benötigt (vgl. Stock 2007: 9–10, 95; Siddiqui & Tiwary 2008: 301–303). Die Funktionsweise von Text Retrieval Systemen ist schematisch in Abbildung 5 skizziert (adaptiert von Jurafsky & Martin 2009: 802).

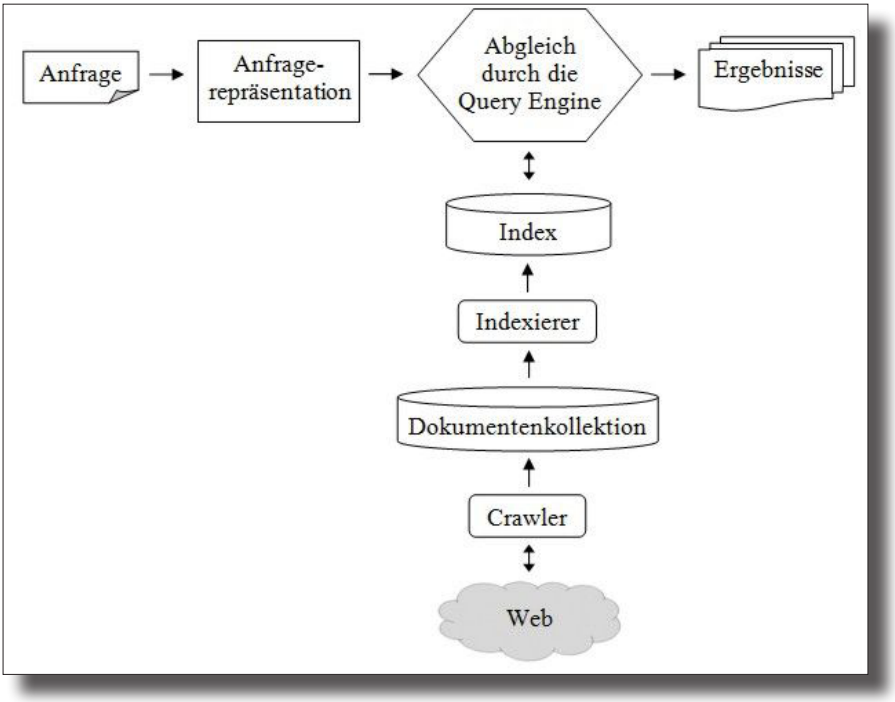

Abb. 5: Funktionsweise von Text Retrieval Systemen im Web

Den Grundstein für ein Text Retrieval System bildet die Datenbank oder Dokumentenkollektion, in die ein Crawler möglichst viele Dokumente bzw. Webseiten aus dem Web einspeist. Ist die Datenbank erstellt, werden die darin gespeicherten Dokumente aufbereitet. Dabei werden wesentliche Merkmale von Dokumenten mittels eines Indexierers in einen Index überführt. Wesentlich ist meist das, was der Benutzer auch auf der Webseite sehen kann. So sind HTML-Tags mit Metakommentaren oder Formatierungen, die der Struktur eines Textes dienen, wie die Position der Absätze oder Kursivsetzungen von Textpassagen, in der Regel nicht relevant; diese Informationen werden daher entfernt. Anschließend wird meist eine Tokenisierung, manchmal auch noch eine Stoppwort-Erkennung und ein Stemming, durchgeführt. Bei einer Tokenisierung wird eine Folge von Zeichen in einzelne Terme gesplittet; ein Term ist dabei meist gleichbedeutend mit einem Wort. Mit der Tokenisierung können nun Wortlisten erstellt werden, die nicht nur speichern, in welchem Dokument welches Wort auftritt, sondern auch noch, wie häufig jedes Wort vorkommt. Da nicht alle Wörter im selben Ausmaß für den Inhalt eines Textes maßgeblich sind, werden manchmal sogenannte Stoppwörter gelöscht. Zu Stoppwörtern gehören Artikel wie *der* oder Präpositionen wie *in*. Das Stemming dient schließlich dazu, Ausdrücke mit verschiedenen Endungen zu einem Wort zusammenzuführen. So etwa handelt es sich bei *Bibliothekar*, *Bibliothekars* und *Bibliothekare* um dieselbe Berufsgruppe. Ohne Stemming würden sie jedoch als drei verschiedene Terme in die Wortliste eingehen. Unterziehen sie sich dem Stemming, werden die Endungen *-s* und *-e* gestrichen und allein der Ausdruck *Bibliothekar* mit einer Häufigkeit von drei in den Index aufgenommen (vgl. Jackson & Moulinier 2002: 10–12; Henrich 2007: 93–98; Stock 2007: 224–225, 232–235, 276–281, 298–299; Levene 2010: 78–82).

Wie in Abbildung 5 weiter zu sehen ist, wird die Suchanfrage, die ein Benutzer stellt, ebenso wie die Dokumentenkollektion aufbereitet und in eine geeignete Form überführt. Nun kann die Anfrage mit der Datenbank abgeglichen und es können all diejenigen Dokumente ausgegeben werden, die für die Anfrage möglichst relevant erscheinen. Da es bei vielen Suchbegriffen meist sehr viele „irgendwie" passende Dokumente im Netz gibt, ist es bedeutsam, die als relevant erachteten Dokumente nicht wahllos, sondern nach Relevanz geordnet dem Benutzer zu präsentieren. Fuhr (2011: 4) beispielsweise behauptet, dass 90 % der Benutzer nur die ersten zehn Treffer in Augenschein nehmen. Zum Ranking von Dokumenten wird die Termhäufigkeit herangezogen, das bedeutet, dass ein Dokument als relevanter erachtet wird, wenn der gewünschte Suchbegriff häufiger auftritt. Andere Kriterien sind der Status der Webseite, also je mehr Links

auf die betreffende Seite verweisen und je häufiger sie von Internetnutzern besucht wird, desto wichtiger ist die Webseite (vgl. Henrich 2007: 357–367, 380–382; Stock 2007: 382–385; Jurafsky & Martin 2009: 117; Levene 2010: 96–106).

5. Potenzial von Anaphern in Suchmaschinen

Eine Anaphernresolution ist besonders dann von Vorteil, wenn das Textverständnis eine Rolle spielt. Eine computergestützte Anaphernresolution wird deswegen vor allem in folgenden Bereichen eingesetzt: Maschinelle Übersetzung, Informationsextraktion, Question Answering sowie Textzusammenfassung. Wenig Beachtung fand bisher die Anwendung im Text Retrieval (vgl. Mitkov 2002: 123–125).

Bei der *maschinellen Übersetzung* bereiten vor allem die Unterschiede zwischen den Sprachen Schwierigkeiten. Beispielsweise sind auf Ebene der Syntax, also wie Elemente einen Satz konstituieren, je nach Sprache verschiedene Realisierungen üblich. Jede Sprache hat außerdem ein eigenes Lexikon, wobei bestimmte Wörter nicht unbedingt Äquivalente in einer anderen Sprache haben müssen. Typische Beispiele betreffen kulinarische Gerichte, wie das *Schnitzel*. Ist jedoch eine grobe Übersetzung ausreichend, so gibt es bereits Beispiele wie das Online-Tool Google Translate (http://translate.google.com, letzter Zugriff: 22.03.2014) (vgl. Jurafsky & Martin 2009: 895–902). Die Anaphernresolution ist in der maschinellen Übersetzung wichtig, da Sprachen sich hinsichtlich der Verwendung spezifischer Anaphern unterscheiden. So beispielsweise wird im Englischen das Pronomen *it* verwendet, um auf *the sun* zu verweisen. Im Deutschen jedoch wird das Pronomen *sie* gebraucht. Ähnlich werden im Spanischen in unmarkierten Sätzen Personalpronomen ausgelassen, beispielsweise in *Tiene un coche*. Bei der Übersetzung ins Englische jedoch muss das Personalpronomen eruiert werden: *She/he has a car.* (vgl. Eberle 2003: 216–217).

Bei der *Informationsextraktion* wird dem Benutzer nicht der gesamte Text zurückgegeben, sondern es werden nur bestimmte Informationen herausgefiltert. Häufig wird die Informationsextraktion im Nachrichtensektor angewandt. Typischerweise werden Personennamen inklusive etwa deren Webseite, E-Mailadresse und Telefonnummer extrahiert oder auch Events mit Details über das Ereignis, dem Ort, dem Datum und der Zeit aus Texten gefiltert. Ein Beispiel ist die Suchmaschine ZoomInfo (http://www.zoominfo.com, letzter Zugriff: 22.03.2014), die allein Informationen über Personen und Firmen extrahiert. Die Anaphernresolution kann hier helfen,

um Einheiten, die koreferentiell sind, zu detektieren und damit die Extraktion zu verbessern. Koreferenz besteht dann, wenn zwei Ausdrücke auf ein und dieselbe Person oder ein und denselben Gegenstand oder Sachverhalt in der Welt verweisen; zum Beispiel handelt es sich gegenwärtig bei *Barack Obama* und *the current president of the USA* um ein und dieselbe Person (vgl. Siddiqui & Tiwary 2008: 337–338, 342; Jurafsky & Martin 2009: 759).

Beim *Question Answering* formuliert ein Benutzer eine Frage und das System beantwortet diese nicht mittels Rückgabe relevanter Dokumente, sondern in Form eines Wortes oder einer kurzen Textpassage, in der die Antwort auf die Frage enthalten sein müsste. Beispielsweise gibt ein Question Answering System auf die Frage des Benutzers „Was ist die Hauptstadt von Österreich?" die Antwort „Wien". Ein Question Answering System im Web ist Ask (http://www.ask.com, letzter Zugriff: 23.03.2014) (vgl. Siddiqui & Tiwary 2008: 358–364; Jurafsky & Martin 2009: 799). Beim Question Answering dient die Anaphernresolution dazu, Koreferenz zwischen Einheiten der Frage und den Dokumenten, in denen die Antwort möglicherweise zu finden ist, herzustellen (vgl. Jurafsky & Martin 2009: 799).

Die *Textzusammenfassung*, wie der Name bereits suggeriert, erstellt automatisch eine kurze Version eines längeren Textes. Aufgrund dieser Zusammenfassung soll der Benutzer entscheiden können, ob der Text für ihn relevant ist oder nicht. Dies ist insofern hilfreich, als dass nicht sofort der ganze Text gelesen werden muss, ohne überhaupt zu wissen, ob der Text überhaupt relevant ist. Bei der Textzusammenfassung werden einzelne Sätze aus einem Dokument ausgewählt. Dabei kann es bei einer fehlenden Anaphernresolution vorkommen, dass Anaphern ohne ihr Antezedens in der Zusammenfassung erscheinen. Dies kann zu einer erschwerten Lesbarkeit oder sogar zu Missverständnissen führen (vgl. Siddiqui & Tiwary 2008: 347–351; Jurafsky & Martin 2009: 822, 836).

Schließlich kann die Anaphernresolution auch im *Text Retrieval* angewandt werden. Stock (2007: 147–150, 295–299) erläutert, dass dies für zwei Bereiche vorteilhaft ist: Die Anaphernauflösung verbessert einerseits die Suche mit Abstandsoperatoren und andererseits beeinflusst sie die Termfrequenz – und damit ganz entscheidend die Suchqualität. Abstandsoperatoren werden eingesetzt, um zwischen zwei Ausdrücken nur eine bestimmte Anzahl an Wörtern zuzulassen, was dazu dient, ein besseres Ergebnis zu erhalten. Sucht man beispielsweise nach *Passau* und *university* und setzt den Abstandsoperator auf zehn, dann wird das Dokument mit folgendem Satz gefunden: *The university that is located in Passau is beautiful.* Das Dokument mit dem Inhalt in Beispiel (14) wird ohne Anaphernresolution jedoch nicht an den Benutzer ausgegeben, da sich zwischen den bei-

den Ausdrücken 15 Wörter befinden. Werden diese zwei Sätze einer Anaphernresolution unterzogen, würde sich die Passage wie in Beispiel (15) verändern und damit das Dokument wieder relevant sein. Manche Suchmaschinen wie Exalead (http://www.exalead.com/search, letzter Zugriff: 19.03.2014) bieten Abstandsoperatoren an (vgl. „Exalead: Web Search Syntax“ 2014). Google scheint den Operator AROUND(n) anzubieten, bei dem Benutzer selbst entscheiden können, wie viele Wörter sich zwischen zwei Ausdrücken befinden dürfen: *university AROUND(1) Passau* liefert Dokumente zurück, bei denen sich zwischen den beiden Ausdrücken ein Wort befindet (vgl. Agarwal 06.02.2012).

(14) The university has about 10,000 students. It is a rather small institution, embedded in the beautiful city Passau.
(15) The university has about 10,000 students. The university is a rather small institution; the university is embedded in the beautiful city Passau.

Die Anaphernresolution ist außerdem für eine bessere Indexierung von Texten hilfreich. Wie in Abschnitt 4 erläutert, wird der Inhalt von Texten in einem Index repräsentiert. Dazu wird gezählt, wie oft ein Wort in einem Dokument auftritt; die Terme werden dann zusammen mit ihrer Häufigkeit in diesem Dokument im Index gespeichert. Werden Anaphern nicht aufgelöst, so verzerrt dies nun die Termfrequenz, da eine Anapher und deren Antezedens nicht auf einen Term reduziert werden können (vgl. Stock 2007: 147–150, 225, 298–299). So würde in der Textpassage in Beispiel (14) ohne Anaphernauflösung der Ausdruck *university* mit der Häufigkeit von eins in den Index eingehen. Mit Anaphernauflösung (Beispiel 15) würde *university* jedoch die Häufigkeit von drei erhalten, was viel besser auch dem Inhalt dieser Passage entspricht. Somit kann die Anaphernresolution die Suche nach Texten entscheidend verbessern.

Obgleich des Nutzens einer Anaphernresolution in Text Retrieval Systemen wurde dazu bislang nur wenig Forschung betrieben. Ganz anders wurde der Einsatz der Anaphernresolution in der maschinellen Übersetzung, der Informationsextraktion, dem Question Answering und der Textzusammenfassung bereits intensiv untersucht. Zur Forschung im Text Retrieval sei vor allem auf Pirkola (1999) hinsichtlich Abstandsoperatoren und auf Liddy (1990) bezüglich der Termfrequenz verwiesen. Beide zeigten entscheidende Verbesserungen, was nicht überraschen sollte, da Anaphern häufig auf Schlüsselbegriffe in Texten verweisen und Anaphern anstatt wortidenter Wiederholungen verwendet werden.

6. Fazit

Werden Theorien und Methoden der Sprachwissenschaft zielgerichtet in computergestützten Anwendungen eingesetzt, so bietet sich ein großes Potenzial, wie deren Ergebnisse verbessert werden können. Hier wurde gezeigt, wie eine präzise Definition und Kategorisierung der Anaphern mit Blick sowohl auf die Sprachwissenschaft als auch auf die Informatik eine bisher unbeachtete Anaphernart, die *non-finite clause* Anaphern, identifizieren können und sich diese bei näherer Analyse sogar als die häufigste Anaphernart überhaupt herauskristallisierte. Diese Ergebnisse bestätigen, dass die Sprachwissenschaft nicht außer Acht gelassen werden darf und künftig mehr Bedeutung in Bereichen der natürlichen Sprachverarbeitung einnehmen sollte.

Danksagung

Dieser Beitrag entstand aus der Doktorarbeit „Anaphora Resolution and Text Retrieval: A Linguistic Analysis of Hypertexts“, die mit dem Förderungspreis 2013 des Vereins zur Förderung der Informationswissenschaft (VFI) ausgezeichnet wurde. Ich möchte mich an dieser Stelle bei der Vergabekommission ganz herzlich für diesen Preis bedanken.

Dr. Helene Schmolz
Lehrstuhl für Englische Sprache und Kultur
Universität Passau, Deutschland
E-Mail: helene.schmolz@uni-passau.de
GND-ID-Nr.: 105358539X

Literatur

Agarwal, Amit (06.02.2012), „A Google Search Operator That You May Not Know About!“, *Digital Inspiration*, http://www.booleanblackbelt.com/2011/06/beyond-boolean-search-proximity-and-weighting/ (letzter Zugriff: 28.03.2014).

Baldwin, Breck (1997), „CogNIAC: High Precision Coreference with Limited Knowledge and Linguistic Resources“, *in* Ruslan Mitkov & Branimir Boguraev, Hg., *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, S. 38–45.

Eberle, Kurt (2003), „Anaphernresolution in flach analysierten Texten für Recherche und Übersetzung“, *in* Ute Seewald-Heeg, Hg., *Sprachtechnologie für die multilinguale Kommunikation. Textproduktion, Recherche, Übersetzung, Lokalisierung*, Sankt Augustin: Gardez!, S. 216–232.

„Exalead: Web Search Syntax“ (2014), http://www.exalead.com/search/web/search-syntax/ (letzter Zugriff: 26.03.2014).

Fuhr, Norbert (2011), „Einführung in Information Retrieval. Skriptum zur Vorlesung im WS 2011/12“, http://www.is.informatik.uni-duisburg.de/courses/ir_ws11/folien/skript_1-6.pdf (letzter Zugriff: 28.03.2014).

Haghighi, Aria & Dan Klein (2009), „Simple Coreference Resolution with Rich Syntactic and Semantic Features“, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, S. 1152–1161.

Henrich, Andreas (2007), *Information Retrieval 1. Kurs im Wintersemester 2007/2008* (Skript zum VHB-Kurs), 2008 publiziert als *Information Retrieval 1. Grundlagen, Modelle und Anwendungen*, Bamberg: Otto-Friedrich-Universität Bamberg, http://www.uni-bamberg.de/minf/ir1-buch/ (letzter Zugriff: 16.03.2014).

Hobbs, Jerry R. (1976), „Pronoun Resolution“ (Forschungsbericht), New York: City University of New York, http://www.isi.edu/~hobbs/PronounResolution.pdf (letzter Zugriff: 04.04.2014).

Jackson, Peter & Isabelle Moulinier (2002), *Natural Language Processing for Online Applications. Text Retrieval, Extraction and Categorization*, Amsterdam – Philadelphia: Benjamins.

Jurafsky, Daniel & James H. Martin (2009), *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2. Aufl.), Upper Saddle River, New Jersey: Pearson.

Lappin, Shalom & Herbert J. Leass (1994), „An Algorithm for Pronominal Anaphora Resolution“, *Computational Linguistics*, 20 (4), S. 535–561.

Levene, Mark (2010), *An Introduction to Search Engines and Web Navigation* (2. Aufl.), Hoboken, NJ: Wiley.

Liddy, Elizabeth DuRoss (1990), „Anaphora in Natural Language Processing and Information Retrieval“, *Information Processing & Management*, 26 (1), S. 39-52.

Mitkov, Ruslan (2002), *Anaphora Resolution*, London et al.: Longman.

Mitkov, Ruslan & Catalina Hallett (2007), „Comparing Pronoun Resolution Algorithms“, *Computational Intelligence*, 23 (2), S. 262–297.

Pirkola, Ari (1999), „Studies on Linguistic Problems and Methods in Text Retrieval. The Effects of Anaphor and Ellipsis Resolution in Proximity Searching, and Translation and Query Structuring Methods in Cross-Language Retrieval“ (Doktorarbeit), Tampere: University of Tampere,

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.138.8052&rep=rep1&type=pdf (letzter Zugriff: 27.03.2014).
Schmolz, Helene & David Coquil (2014), „Anaphernresolution im Text Retrieval – Ein sprachwissenschaftlich-informationstechnologischer Ansatz zur qualitativen Verbesserung von Suchmaschinen“, *in* Institut für interdisziplinäre Medienforschung, *Suchmaschinen*, Berlin: Logos, S. 63–84.
Schmolz, Helene, Mario Döller & David Coquil (2012), „In-Depth Analysis of Anaphora Resolution Requirements“, *Proceedings, TIR-Workshop, DEXA-Konferenz (Wien)*, Los Alamitos et al.: IEEE, S. 174–179.
Siddiqui, Tanveer & Uma S. Tiwary (2008), *Natural Language Processing and Information Retrieval*, New Delhi: Oxford University Press.
Soon, Wee Meng, Hwee Tou Ng & Daniel Chung Yong Lim (2001), „A Machine Learning Approach to Coreference Resolution of Noun Phrases“, *Computational Linguistics*, 27 (4), S. 521–544.
Stock, Wolfgang (2007), *Information Retrieval. Informationen suchen und finden*, München – Wien: Oldenbourg.
Stoyanov, Veselin et al. (2010), „Coreference Resolution with Reconcile“, *Proceedings of the ACL 2010 Conference Short Papers*, S. 156–161.
Strube, Michael (2010), „Anaphernresolution“, *in* Kai-Uwe Carstensen et al., Hg., *Computerlinguistik und Sprachtechnologie. Eine Einführung* (3. Aufl.), Heidelberg: Spektrum, S. 399–409.
Uryupina, Olga (2010), „Corry: A System for Coreference Resolution“, *Proceedings of 5th International Workshop on Semantic Evaluation*, S. 100–103.
Versley, Yannick et al. (2008), „BART: A Modular Toolkit for Coreference Resolution“, *Proceedings of the ACL-08: HLT Demo Session*, S. 9–12.