MICHAEL STRUBE/MARGOT MIESKES/CHRISTOPH MÜLLER

Gesprächsprotokolle auf Knopfdruck: Die automatische Zusammenfassung von gesprochenen Dialogen

Abstract

Dieser Beitrag beschreibt computerlinguistische Arbeiten zur automatischen Zusammenfassung gesprochener Dialoge. Der Beitrag geht sowohl auf die notwendige Vorverarbeitung als auch auf die eigentliche Zusammenfassung durch automatische Erkennung von Themengrenzen und Extraktion relevanter Äußerungen ein. Ein weiterer Schwerpunkt liegt in der Beschreibung von Arbeiten zur automatischen Anaphernresolution in gesprochener Sprache. Der Beitrag betont vor allem die Rolle und Bedeutung von annotierten Korpora für die computerlinguistische Forschung und Entwicklung.

1. Einleitung

Täglich finden unzählige Besprechungen, Verhandlungen und Diskussionen statt. Oft schreibt einer der Teilnehmer fleißig Notizen, um später ein Protokoll anzufertigen – eine meist ungeliebte Aufgabe. Im von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekt DIANA-Summ (*Dialog ANAphors and Summarization*) soll ein Computersystem entwickelt werden, das gesprochene Dialoge automatisch zusammenfasst. Langfristiges Ziel des Projektes ist es, die automatische Erzeugung von Gesprächsprotokollen zu ermöglichen. Das Verfahren, das dabei zur Anwendung kommt, basiert darauf, für jedes im Gespräch behandelte Thema (*topic*) relevante Äußerungen zu extrahieren und zu einer Zusammenfassung zu kombinieren.

Spontane gesprochene Sprache unterscheidet sich in mehrfacher Hinsicht von geplanter, geschriebener Sprache. Zum einen enthält sie viele abgebrochene Äußerungen, Wiederholungen, Selbstkorrekturen und Pausen, wobei letztere häufig mit Platzhaltern wie 'Ähh' gefüllt sind. Sind mehrere Sprecher beteiligt, kommt als weitere Komplikation hinzu, dass diese auch gleichzeitig sprechen können. Dies alles macht es erforderlich, dass vor der Erstellung einer Zusammenfassung eine Reihe von Vorverarbeitungsschritten durchgeführt wird, die diese Phänomene erkennen und ggf. entfernen. Eine weitere Besonderheit gesprochener Sprache ist der erheblich höhere Anteil an Pronomen wie (im Deutschen) *es* und *das* oder (im Englischen) *it* und *this* bzw. *that*. Eine der Hypothesen des Projektes ist, dass die Auflösung dieser Pronomen

(d. h. die Bestimmung ihrer jeweiligen Antezedenten) sowohl die Identifikation von relevanten Äußerungen erleichtern als auch die Verständlichkeit der so erstellten Zusammenfassungen erhöhen sollte.

Die folgenden Kapitel geben einen Überblick über das Projekt DIANA-Summ (Stand April 2006). Nach einer Beschreibung der verwendeten Datengrundlage (Kap. 2) wird zunächst auf methodische Aspekte des Projektes eingegangen (Kap. 3). Dabei steht insbesondere die Rolle des Korpus als empirische Grundlage für Systementwicklung und Evaluierung im Mittelpunkt. Außerdem wird das Konzept der Korpusannotation eingeführt. Datengewinnung durch Korpusannotation sowie die Sicherung der Qualität der so gewonnenen Daten werden ebenfalls erläutert. Im Anschluss daran werden die einzelnen Schritte der Vorverarbeitungskette (Kap. 4) beschrieben. Die drei letzten Kapitel gehen dann detaillierter auf die Teilaufgaben Anaphernresolution (Kap. 5), Erkennung von Themengrenzen (Kap. 6) und Zusammenfassung (Kap. 7) ein.

2. Datengrundlage

Im Rahmen des DIANA-Summ-Projektes soll keine eigene Sammlung von Rohdaten durchgeführt werden, da derartige Datenerhebungen mit erheblichem technischen und organisatorischen Aufwand verbunden, für die eigentliche Zielsetzung des Projektes aber nicht von besonderem Interesse sind. Als Datengrundlage wird deshalb das englischsprachige *ICSI Meeting Corpus* (Janin et al., 2003) verwendet. Bei diesem Korpus handelt es sich um Aufnahmen von 75 Diskussionen (gesamt ca. 72 Stunden). Jede Diskussion hat durchschnittlich 6 Teilnehmer, wobei nicht alle Teilnehmer Englische Muttersprachler sind. Themen umfassen u. a. Computerlinguistik, Sprachtheorie, Spracherkennung und Netzwerktechnologie. Zusätzlich zu den eigentlichen Sprecheräußerungen enthält das Korpus auch explizite Markierungen von abgebrochenen Äußerungen bzw. Wörtern (sog. *disfluencies*, dt. Häitationen), sowie Meta-Informationen über nicht-sprachliche Ereignisse (Atemgeräusche, Rauschen etc.). In struktureller Hinsicht besteht das Korpus aus einer Reihe von sog. Segmenten. Jedes Segment ist mit genau einem Sprecher assoziiert, wobei längere Äußerungen desselben Sprechers auch auf mehrere Segmente verteilt sein können. Vergleiche das folgende Beispiel (Nummerierung nicht im Original).

1. FE016 uh, echo cancellation to get basically back down to the individual
2. FE016 speaker utterances would be probably all that we would need to be able to do good recognition on the – on the close-talking mikes.
3. ME011 Um, why don't you, if you have a hard copy, why don't you email it
4. MN014 So these –
5. FE016 So, that's about the summary –
6. ME011 to the list.
7. FE016 But this is – Morgan has this paper. I mean he – he – it – it's that paper. Yeah, yeah.

8. MN014 Yeah, yeah.
9. ME013 Yeah, so it's the same thing? It's the same thing I mailed to every –
10. ME011 Oh it's in the paper.
11. ME011 O_K.
12. ME013 everybody that w- where it was, yeah.
13. FE016 So, we basically, um,
14. ME011 O_K then, it's already been mailed.

Es fällt sofort der Anteil an überlappender Sprache auf, der sich in den Daten in Form von diskontinuierlichen Segmenten (z. B. 3–6 bzw. 9–12) manifestiert. Auch ist erkennbar, dass Segmentgrenzen nicht unbedingt mit linguistisch motivierten Phrasen- oder Satzgrenzen zusammenfallen müssen (z. B. 1–2). Unterbrechungen bzw. abgebrochene Äußerungen sind durch Bindestriche angezeigt (z. B. 7 und 9).

Das *ICSI Meeting Corpus* liegt vollständig manuell transkribiert und im XML-Format vor. Da automatische Spracherkenner spontane Diskussionen mit mehreren Teilnehmern noch nicht in befriedigender Qualität verarbeiten können, wird für DIANA-Summ eine idealisierte Eingabe in Form der manuell erstellten Transkriptionen angenommen.

3. Methoden

In der Korpuslinguistik werden Textkorpora schon lange als maschinenlesbare Sammlungen natürlicher Sprache verwendet. Sie dienen dort als empirische Grundlage für Häufigkeits-, Kollokations- oder andere Analysen. Im Gegensatz dazu geht die Verwendung von Korpora in der Computerlinguistik über die reine Deskription hinaus, denn ihre Aufgabe ist die Entwicklung von Computersystemen zur automatischen Text- oder Sprachverarbeitung. Im Projekt DIANA-Summ sind dies in erster Linie die automatische extraktive Zusammenfassung und die Anaphernresolution. Ein weit verbreiteter methodischer Ansatz für die Entwicklung derartiger Systeme, der auch im Rahmen von DIANA-Summ Anwendung findet, ist der Einsatz von sog. überwachtem maschinellen Lernen (*supervised machine learning* (ML)) (Mitchell, 1997). Bei dieser Methode stellen annotierte Korpora eine unverzichtbare empirische Grundlage dar. Die Korpora werden mit Hilfe spezieller ML-Programme analysiert. Die Programme erzeugen statistische Modelle der in den Korpora vorgefundenen Phänomene, welche dann auf neue Korpora angewendet werden können. Die ML-Programme können allerdings nur solche Phänomene des analysierten Korpus in ihrem Modell berücksichtigen, welche explizit in diesem enthalten sind. Um z. B. ein Modell für die automatische Erkennung von Häitationen zu erzeugen (zu *trainieren*), muss das analysierte Trainingskorpus für jedes Wort Informationen darüber enthalten, ob es Teil einer Häitation ist. Das gleiche gilt für das Trainieren von Modellen zur Anaphernresolution: Jeder anaphorische Ausdruck im Trainingskorpus muss mit seinem korrekten Antezedenten verbunden sein, damit ein Modell trainiert werden

kann, das für jede Anapher den *wahrscheinlichsten* Antezedenten identifizieren kann.

Der Vorgang und das Ergebnis der manuellen Hinzufügung dieser Art von Information zu einem existierenden Textkorpus wird *Annotation* genannt. Außer zum Training von statistischen Modellen für die automatische Text- oder Sprachverarbeitung werden annotierte Korpora auch zu deren Evaluierung benötigt. Dabei wird die Leistungsfähigkeit eines Systems bestimmt, indem seine (potentiell falsche) Ausgabe mit der zuvor manuell erstellten, korrekten Annotation im Korpus verglichen wird. Bei der Evaluierung, welche ein wichtiger Schritt in der Entwicklung computerlinguistischer Systeme ist, kommen verschiedene quantitative Maße zum Einsatz, darunter v. a. Präzision (*precision*) und Vollständigkeit (*recall*). Präzision wird dabei berechnet aus „Anzahl der korrekten Antworten des Systems“ geteilt durch „Anzahl aller Antworten des Systems“. Die Vollständigkeit berechnet sich aus „Anzahl der korrekten Antworten des Systems“ geteilt durch „Anzahl der möglichen korrekten Antworten“. Das F-Maß (*f-measure*) setzt Präzision und Vollständigkeit ins Verhältnis zueinander.

Linguistische Annotation ist eine mühsame und zeitraubende Aufgabe. Da sie größtenteils manuell durchgeführt wird, ist sie auch mit erheblichen Kosten verbunden. Die Annotierer verwenden in der Regel spezialisierte Annotations-Tools (wie z. B. MMAX2 (<http://mmax.eml-research.de>)). Je nach der Schwierigkeit der Aufgabe (d. h. des zu annotierenden Phänomens) kann die manuelle Annotation erhebliche Interpretation von seiten der Annotierer erfordern, so dass verschiedene Annotierer selten völlig identische Annotationen produzieren. Der Grad der Übereinstimmung zwischen zwei (oder mehr) Annotierern für dieselbe Aufgabe (*inter-annotator agreement*) ist daher ein wichtiger Indikator für die Schwierigkeit bzw. die grundsätzliche Lösbarkeit der Annotationsaufgabe. Die Übereinstimmung wird in der Regel in Form der Kappa-Statistik berechnet (Carletta, 1996). Das Besondere an diesem Maß ist, dass es berücksichtigt, dass ein bestimmter Grad an Übereinstimmung auch durch Zufall zustande kommen kann. In der Regel bewegt sich Kappa zwischen -1 und 1 . Ein niedriger Kappa-Wert für eine bestimmte Annotationsaufgabe kann ein Hinweis darauf sein, dass das zu annotierende Phänomen an sich unscharf, ambig oder vage ist und keine eindeutige Lösung besitzt. Im Allgemeinen wird ein Kappa-Wert von $0,67$ als gerade noch akzeptabel angesehen. Liegt der Wert darunter, dann sind die durch die Annotation gewonnenen Daten nur von sehr eingeschränkter Nützlichkeit. Angesichts der zentralen Bedeutung der manuellen Annotation im Projekt DIANA-Summ ist es daher notwendig, die Datenqualität sowohl für die Vorverarbeitung (Kap. 4) als auch für die eigentlichen Aufgaben (Kap. 5, 6 und 7) durch Kontrolle des Kappa-Wertes zu sichern.

4. Vorverarbeitungsschritte

Um die transkribierten Dialoge zusammenfassen zu können, sind mehrere Vorverarbeitungsschritte nötig. Einige davon werden wir im Folgenden näher betrachten. Eine schematische Darstellung der Vorverarbeitung findet sich in Abbildung 1, wobei zu bemerken ist, dass wir nicht alle Vorverarbeitungsschritte selbst durchführen werden. Die so genannten Dialogakte etwa, die jede Aussage mit Information bezüglich der Intention versehen (Bitte, Frage, Aussage etc), wurden uns von einer anderen Forschungsgruppe zur Verfügung gestellt.

Abb. 1

4.1 Part-of-Speech Tagging

Part-of-Speech (POS) Tagging bezeichnet den Vorgang der Anreicherung/Annotation der Daten mit Wortarten. Da es ausgereifte Programme gibt, die die Annotation mit Wortarten automatisch vornehmen, setzten wir solche ein, die verschiedene Ansätze abdecken: den *Brill Tagger* (TBL; Brill 1994), *TnT* (Brants, 2000), zwei *Tagger* aus der *Stanford NLP Library* (*Left3* und *Bidirect*) (Toutanova et al., 2003, Toutanova und Manning, 2000). Diese Systeme sind

auf den *Wall Street Journal* Daten der *Penn Treebank* (Marcus et al., 1993) trainiert, also auf Zeitungstexten, die sich stark von gesprochener Sprache unterscheiden. Durch einen Mehrheitsentscheid (*Maj*) über allen vier Systemen hofften wir, die ursprüngliche Fehlerrate zu senken, die sowohl durch die Anwendung auf Dialoge als auch durch die unterschiedlichen Herangehensweisen der verschiedenen *Tagger* entstehen.

Dieser Mehrheitsentscheid wurde Annotierern als Grundlage für eine manuelle Korrektur der automatischen Annotation übergeben. Dies diente zum einen zur Evaluierung der ursprünglichen Annotation, zum anderen aufgrund der subjektiv relativ schlechten Ergebnisse der automatischen *Tagger* für deren Neutrainieren. In zwölf der 75 *Meetings* wurden die *POS-Tags* manuell von allen drei Annotierern korrigiert. In 25 weiteren wurden die *POS-Tags* von einzelnen Annotierern korrigiert. Die Übereinstimmung zwischen den Annotierern wurde mit Hilfe der Kappa-Statistik bestimmt und ist mit $k = 0,96$ sehr hoch. Dies zeigt, dass die Aufgabe der manuellen Korrektur schon vorgefertigter *POS-Tags* sehr zuverlässig möglich ist. Zudem ist dieses Vorgehen deutlich schneller, als eine rein manuelle Annotation.

Aus den manuell korrigierten Daten wurden zwei Testkorpora gewählt. Test1 besteht aus einem Teil der zwölf *Meetings*, die von allen drei Annotierern korrigiert wurden, Test2 besteht aus *Meetings*, die nur von einzelnen Annotierern korrigiert wurden.

Nach der manuellen Annotation waren wir in der Lage, die ursprüngliche Annotation auszuwerten. Die Ergebnisse sind in Tabelle 1 dargestellt.

Data	TBL	TnT	Left3	Bidirect	Maj
Test1	11,3	11,2	11,1	11,1	10,5
Test2	13,8	14,3	13,6	13,2	13,4

Tab. 1

Da diese Ergebnisse weit von den Ergebnissen entfernt sind, die in der Literatur für die *Tagger* angegeben werden (2,5 % bis 3 % Fehlerrate) wurden die manuell korrigierten *POS-Tags* dazu genutzt, die vier automatischen Systeme neu zu trainieren. Dazu wurden die vorhandenen Daten in 6 verschiedene Teilmengen unterteilt, die immer mehr Daten enthielten. Die Ergebnisse des Neutrainierens sind in Tabelle 2 dargestellt, ebenso wie die jeweilige Datenmenge.

Das Neutrainieren bewirkt eine deutliche Verbesserung. Alle vier *Tagger* erreichen Fehlerraten von 3,5 %–4 %, wobei insbesondere bei *TBL* vom ersten zum sechsten Schritt eine deutliche Verbesserung zu sehen ist. Im Durchschnitt liegt die Verbesserung über die Teilmengen bei ca. 1 %–1,5 %, wobei bei allen *Taggern* der Gewinn durch die Erhöhung der Trainingsmenge nach dem vierten oder fünften Schritt nur noch gering ausfällt.

Set Wörter	Set1 124K	Set2 162K	Set3 197K	Set4 221K	Set5 253K	Set6 283K
TnT						
Test1	3,4	3,4	3,3	3,3	3,4	3,4
Test2	5,4	5,1	4,9	4,6	4,5	4,5
TBL						
Test1	3,9	3,5	3,5	3,5	3,6	3,5
Test2	8,4	5,5	5,0	4,7	4,4	4,4
Left3						
Test1	3,2	3,0	3,0	3,2	3,2	3,2
Test2	5,2	4,7	4,5	4,3	4,2	4,1
Bidirect						
Test1	3,2	3,0	3,0	3,2	3,2	3,2
Test2	5,2	4,7	4,5	4,3	4,2	4,1

Tab. 2

Auch mit den neutrainierten Systemen wurde ein Mehrheitsentscheid durchgeführt, allerdings wurde der zweite Stanford-Tagger (*Bidirect*) aufgrund der Ähnlichkeit mit *Left3* nicht weiter berücksichtigt. Das Ergebnis des Mehrheitsentscheids ist in Tabelle 3 illustriert:

	Setup1	Setup2	Setup3	Setup4	Setup5	Setup6
Test1	3,0	2,9	2,9	3,0	3,0	2,9
Test2	5,1	4,7	4,4	4,1	4,0	3,9

Tab. 3

Auch hier ist eine deutliche Verbesserung im Vergleich zu den ursprünglichen Annotationen feststellbar. Allerdings ist die Verbesserung nur bis zum vierten Schritt deutlich und schwächt sich danach ab. Man kann daraus schließen, daß der erhöhte Aufwand durch das manuelle Annotieren ab dem vierten bis fünften Schritt nicht mehr zu einer deutlichen Verbesserung der Ergebnisse führt.

Der einfache Mehrheitsentscheid bietet also eine computationell unaufwendige Methode, um die Ergebnisse weiter zu verbessern. Damit liegt die Fehlerrate beim *Taggen* von transkribierter, gesprochener Sprache ähnlich niedrig wie die von geschriebener Sprache. Eine detailliertere Ausführung zu diesem Thema findet sich in Mieskes und Strube (2006).

4.2 Häitationen und Unterbrechungen

Häitationen, Unterbrechungen und Selbstkorrekturen sind ein Phänomen der gesprochenen Sprache. In geschriebenen Texten kommen sie nicht vor und nur äußerst selten bei gelesenen Texten wie beispielsweise Nachrichtentexten. Dabei unterscheidet man vier Hauptformen: Pausen (nichtlexikalische (NLFP) und lexikalische (LFP)), Wiederholungen (*repet*), Fehlstarts (abgebrochene Wörter (*abw*) und abgebrochene Äußerungen (*abutt*)) und Selbstkorrekturen (*repai*). Hierbei werden nur sogenannte „gefüllte Pausen“ berücksichtigt, da die leeren Pausen (Stille) nur schwer von anderen Pausen (Sprecherwechsel etc.) zu unterscheiden sind. Beispiele für die genannten Häitationen sind:

NLFP: FE016 **uh**, echo cancellation to get basically back down to the individual

LFP: FE016 ... **I mean** he – he – it – it's that paper.

repet: FE016 ... I mean **he – he** – it – it's that paper.

repai: FE016 **But this is** – Morgan has this paper.

abw: ME013 everybody that **w-** where it was, yeah.

abutt: FE016 **So, that's about the summary** –

Auch dieses Phänomen wurde in den Daten manuell annotiert. Werden nur Elemente berücksichtigt, die von beiden Annotierern markiert wurden, ist die Übereinstimmung $k = 0,95$. Bei den einzelnen Kategorien sind die Unterschiede sehr deutlich, wie Tabelle 4 zeigt, wobei die zweite Spalte (inkl. Elemente) alle markierten Elemente berücksichtigt und die dritte Spalte (gem. Elemente) nur jene, die von beiden Annotierern markiert wurden.

Art	inkl. Elemente	gem. Elemente
NLFP	0,52	100,0
LFP	0,47	0,98
Repet	0,63	0,98
Repai	0,24	0,78
Abw	0,54	0,96
Abut	0,28	0,85

Tab. 4

Wie man aus Tabelle 4 ersehen kann, sind besonders Reparaturen, Wiederholungen, aber auch lexikalische Pausen problematisch. Aus den Daten, die von beiden Annotierern markiert wurden, kann man jedoch erkennen, dass es eine grosse Übereinstimmung gibt. Dies ist gut genug, um eine automatische Erkennung von Häitationen zu entwickeln. Eine erste Mustererkennung, die

auf einfachen Buchstaben- und Wortabgleichungen basiert, erlaubt es, die nichtlexikalischen Pausen und wortwörtliche Wiederholungen zu bestimmen. Für die komplexeren Phänomene wurden Methoden des maschinellen Lernens eingesetzt. Die besten Ergebnisse mit mehr *Features*, als nur den drei vorangehenden *POS-Tags* sind in Tabelle 5 aufgeführt.

	total	richtig	falsch	Präz.	Vollst.	F-Maß
NLFP	8110	7404	2506	74,71	91,92	82,18
LFP	8489	6881	1737	79,84	81,06	80,45
abutt	4568	2198	1541	58,70	47,94	52,78
abw	2531	2275	989	69,70	89,89	78,52
repai	6620	3949	2494	61,29	59,65	60,46
repet	5156	3677	1347	73,19	71,19	72,18

Tab. 5

4.3 Lemmatisierung

Auf Basis der *POS*-Annotation lässt sich ein weiterer Vorverarbeitungsschritt automatisieren, der für die Zusammenfassung relevant sein wird, die sogenannte Lemmatisierung. Hierbei werden die Grundformen der Wörter ermittelt. Beispielsweise sind *is* und *was* beide Formen des Verbs *be*. *houses* ist die Pluralform von *house*. *be* und *house* sind jeweils die Lemmata von *is* und *was*, resp. *houses*. Für die Bestimmung der Lemmata verwendeten wir den *Tree-Tagger* (Schmid, 1997).

5. Anaphernerkennung und -resolution

In der Anaphernresolution im Projekt DIANA-Summ wird ein besonderes Augenmerk auf das Personalpronomen *it* und die Demonstrativpronomina *this* und *that* gelegt. Die Präferenz für diese Pronomen liegt neben ihrer Häufigkeit auch in der Anwendung begründet, in welche die Anaphernresolution eingebettet und anhand welcher sie zu evaluieren ist. Im Kontext einer extraktiven Zusammenfassung ist es nahe liegend, einzelne Äußerungen durch Resolution der in ihnen enthaltenen Pronomen und durch Ersetzung derselben durch ihre lexikalischen Antezedenten semantisch explizit und damit von ihrem Kontext unabhängig zu machen. Dies hat positive Auswirkungen sowohl auf die Extraktion als auch auf die Lesbarkeit der erzeugten Zusammenfassung.

Nicht alle Vorkommen von *it*, *this* und *that* sind indes resolvierbare Anaphern. Vor der eigentlichen Anaphernresolution muss daher eine Desambiguierung bzw. Vorauswahl erfolgen. Für *this* und *that* z. B. müssen pronominale Verwendungen von solchen unterschieden werden, in denen die Aus-

drücke als Determinierer einer NP fungieren. *That* kann außerdem auch Relativpronomen oder Komplementierer (entsprechend Dt. *dass*) sein. Die automatische Unterscheidung all dieser Fälle wird durch die in Kapitel 4 beschriebene Wortartenklassifizierung geleistet. Die Identifikation der nicht-referentiellen und somit nicht resolvierbaren Fälle des Pronomens *it* ist hingegen ungleich schwieriger. Der folgende Abschnitt gibt lediglich einen Überblick. Eine detailliertere Darstellung (inklusive Beispielen) findet sich in Müller (2006).

5.1 Identifikation von nicht-referentiellem *it*

5.1.1 Typen von *it*

In Müller (2006) werden fünf Typen von *it* unterschieden, welche in morpho-syntaktischer Hinsicht alle als Pronomen zu klassifizieren sind. Allerdings sind nur zwei der fünf Typen referentielle Pronomen, und lediglich ein Typ stellt eine resolvierbare Anapher dar.

normal:	Anaphorisches Pronomen mit identifizierbarem Antezedenten (referentiell, resolvierbar)
vage:	Anaphorisches Pronomen ohne identifizierbaren Antezedenten (referentiell, nicht resolvierbar)
verworfen:	Pronomen in abgebrochener Äußerung (nicht-referentiell, nicht resolvierbar)
Extraposition:	Pronomen als syntaktischer Platzhalter für nach rechts verschobenes Satzglied (nicht-referentiell, nicht resolvierbar)
Prop-it	(Quirk Pronomen als semantisch leere Besetzung der obligatorischen et al, 1991): Subjekt-Position (nicht-referentiell, nicht resolvierbar)

Als Vorstufe der Resolution des Pronomens *it* ist es daher notwendig, nicht-referentielle Fälle automatisch zu identifizieren und auszufiltern.

5.1.2 Manuelle Annotation von *it*

Zwei Annotierer erhielten die Aufgabe, unabhängig voneinander alle Vorkommen von *it* in einen Sub-Korpus von fünf Dialogen als einen der fünf Typen oder als ‚Sonstiges‘ zu klassifizieren. Die Annotation wurde von zwei Annotierern redundant ausgeführt, um die Übereinstimmung in Form von Kappa ermitteln zu können. Die Resultate dieser ersten Annotation finden sich in Tabelle 6.

Die Kappa-Werte für alle Kategorien mit Ausnahme von ‚verworfen‘ liegen teilweise deutlich unter dem geforderten Wert von 0,67 (vgl. Kap. 3). Ein erstes wichtiges Resultat von Müller (2006) ist daher, dass die differenzierte (d. h. fünffache) Klassifikation der verschiedenen Typen von *it* nicht unproblematisch zu sein scheint, da sie häufig nur mit einer Genauigkeit erfolgt, die an Zufall grenzt. Durch Zusammenlegung der Kategorien ‚normal‘ und

‚vage‘ zu ‚referentiell‘ auf der einen und ‚verworfen‘, ‚Extraposition‘ und ‚Prop-it‘ zu ‚nicht-referentiell‘ auf der anderen Seite kann die Übereinstimmung jedoch deutlich verbessert werden. Die Kappa-Werte für die resultierenden Super-Kategorien ‚referentiell‘ bzw. ‚nicht-referentiell‘ betragen 0,72 bzw. 0,74, so dass die Verwendbarkeit der erhobenen Daten für die weniger differenzierte zweifache Klassifikation gewährleistet ist.

	normal	vage	verwrf.	Extrap.	Prop-it	Sonst.	Ges.
Bed017	0,65	0,33	0,94	0,27	0,54	0,42	0,62
Bmr001	0,69	0,21	0,92	0,48	0,33	– 0,01	0,63
Bns003	0,59	0,18	0,75	0,55	0,21	0,32	0,55
Bro004	0,65	– 0,05	0,86	0,75	0,59	– 0,01	0,65
Bro005	0,57	– 0,03	0,84	0,58	0,36	0,23	0,58
Gesamt	0,64	0,11	0,86	0,56	0,43	0,2	0,61

Tab. 6

Als letzten Schritt der Annotation hatten die Annotierer die Aufgabe, aus ihren teilweise divergierenden Annotationen durch Diskussion von strittigen Fällen eine Referenz-Annotation zu erstellen, welche die Datengrundlage für die im folgenden beschriebene automatische Klassifikation darstellt.

5.1.3 Automatische Klassifikation von *it*

Ziel der automatischen Klassifikation ist es, mit Hilfe von Methoden des maschinellen Lernens (ML) einzelne Vorkommen von *it* als entweder ‚referentiell‘ oder ‚nicht-referentiell‘ zu erkennen. Jedes Vorkommen von *it* wird dafür in Form einer Liste von Merkmalen (*feature vector*) repräsentiert. Diese Merkmale beschreiben den Kontext (im weitesten Sinne), in welchem das *it* vorkommt. Dies sind z. B. syntaktische Konstruktion, regierendes Verb, Distanz (in Wörtern) zu bestimmten grammatischen Funktionswörtern, etc. Jede Merkmalsliste enthält außerdem ein Zielkonzept, d. h. sie ist als ‚referentiell‘ oder ‚nicht-referentiell‘ markiert.

Ein Teil der so gewonnenen Daten wird dann als Trainingsdaten für die ML-Programme genutzt, welche sie analysieren und ein statistisches Modell erzeugen. Dieses Modell kann dann verwendet werden, um neue, nicht im Training verwendete Daten automatisch zu klassifizieren. Die Korrektheit einer jeden Klassifikation kann überprüft werden, indem die automatisch vergebene Klasse mit der während der manuellen Annotation vergebenen verglichen wird. Das beste im Rahmen von Müller (2006) erzielte Ergebnis belief sich auf

Präzision = 80,0%, Vollständigkeit = 60,9% und F-Maß = 69,2%.

Dieses Ergebnis, insbesondere die hohe Präzision, erlaubt es, die hier entwickelte Komponente zur Identifikation von nicht-referentiellem *it* als Filter für die Anaphernresolution einzusetzen.

5.2 Resolution von anaphorischen Pronomen

Während die automatische Anaphernresolution für geschriebene Texte (v. a. Zeitungstexte) bereits eine etablierte Teildisziplin innerhalb der Computerlinguistik ist, steht diese Aufgabe im Bereich der (transkribierten) gesprochenen Sprache noch ziemlich am Anfang. Einer der Hauptgründe dafür ist sicherlich die weitaus größere Schwierigkeit der Aufgabe. Spontane gesprochene Umgangssprache weist einen weitaus höheren Anteil an vagen oder ambigen Referenzausdrücken auf als geschriebener Text. Die in Kapitel 5.1.2 erwähnte geringe Übereinstimmung zwischen Annotierern bei der Klassifikation von normalen und vagen Fällen von *it* kann als Beleg dafür gelten. Im Vergleich zu dieser Aufgabe ist allerdings die Identifikation des korrekten Antezedenten eines Pronomen nochmals erheblich schwieriger. Eine Kontrolle der Annotierer-Übereinstimmung ist demnach auch hier vonnöten.

Die im Rahmen von DIANA-Summ durchzuführende manuelle Koreferenzannotation ist zum aktuellen Zeitpunkt (April 2006) noch nicht abge-

Abb. 2

schlossen, so dass noch keine Zahlen für die Bewertung der Übereinstimmung vorliegen. Als erster Eindruck kann jedoch auch ein qualitativer Vergleich der von den verschiedenen Annotierern gebildeten Koreferenzklassen dienen. Koreferenzklassen kann man am besten in Form von Koreferenzketten visualisieren, indem man alle Elemente einer Klasse in der Reihenfolge ihres Erscheinens im Text verbindet. Diese Form der Visualisierung wird auch vom Annotationstool MMAX2 verwendet. Der *Screenshot* in Abbildung 2 zeigt die graphische Oberfläche des Tools mit einem selektierten Demonstrativpronomen (*this*) an der Cursorposition (Pfeil).

Die unterschiedlich gefärbten Linien (hier in verschiedenen Grautönen wiedergegeben) visualisieren die vier verschiedenen Koreferenzketten, denen das selektierte Pronomen von vier Annotierern unabhängig voneinander zugeordnet wurde. Während dies sicherlich ein Extrembeispiel darstellt, gibt es doch einen guten Eindruck von der Schwierigkeit der Koreferenzannotation in gesprochenen Dialogen.

6. Themengrenzen

Ein wichtiger Schritt für die Zusammenfassung ist die Abgrenzung der verschiedenen Themen, die im Verlauf der Gespräche auftauchen. Da die Gespräche sehr lang sind, behandeln sie häufig mehr als nur ein Thema. In manuellen Protokollen ist es ebenfalls üblich, das Protokoll in thematische Abschnitte zu unterteilen.

In der Vergangenheit wurden verschiedene Methoden zur automatischen Bestimmung der Themengrenzen entwickelt, denen häufig Wortfrequenzen oder statistische Modelle zu Grunde liegen. Neuere Methoden basieren auf lexikalischen Ketten, die zusammengehörende Wörter verbinden. Bisher werden Wörter als zusammengehörend betrachtet, wenn ihnen gleiche Lemmata (vgl. Kapitel 4.3) zu Grunde liegen (beispielsweise *house* und *houses*). Werden lexikalische Ketten über mehrere Sätze hinweg nicht fortgesetzt, so ist das ein Indiz für eine Themengrenze (Galley, 2003).

Auch für die Auswertung automatischer Themengrenzen sind manuelle Annotationen nötig. Da es sich nicht um eine Klassifikations- sondern um eine Segmentierungsaufgabe handelt, müssen die Evaluierungsmethoden berücksichtigen, dass die Grenzen nicht unbedingt exakt übereinstimmen müssen. Evaluierungsmethoden, die diesem Phänomen Rechnung tragen, sind: *Pk* (Beeferman et al., 1999) und *WindowDiff* (Pevzner und Hearst, 2002). Beide Maße geben an, wie groß die Wahrscheinlichkeit ist, eine Grenze falsch zu ziehen. Kleine Zahlen stehen für gute Ergebnisse.

Bei der manuellen Annotation gibt es mehrere Schritte. Zunächst werden nur vorläufige Grenzen gezogen. Diese werden in einem zweiten Schritt korrigiert und zum Teil wieder entfernt. In einem dritten Schritt werden die Grenzen unterschieden in Haupt- und Unterthemen bzw. „*Small Talk*“. *Small Talk* ist für die spätere Zusammenfassung irrelevant und sollte deshalb gefiltert

werden. Bei den hier vorliegenden *Meetings* gibt es viele, in denen es für die menschlichen Annotierer schwierig ist, eindeutige Themengrenzen zu ziehen. Die Ergebnisse liegen je nach Dialog für die manuellen Annotierer zwischen 20,0 und 7,0 (Angabe nach Pk-Maß).

Das erfolgreichste System (im Folgenden *g03* aus Galley et al. (2003)) arbeitet mit lexikalischen Ketten, die auf reiner Wortgleichheit basieren. Dieses System wurde zusammen mit üblichen Baselines getestet. Die Ergebnisse finden sich in Tabelle 7.

	random	even	all	g03
Pk	48,9	55,1	82,3	20,4
WD	53,5	61,3	99,9	24,2

Die drei *Baselines* sind *Random*, wobei eine bestimmte Anzahl an Grenzen (in diesem Fall 25) zufällig im Text gesetzt werden, *Even*, wobei diese 25 Grenzen gleichmäßig über den Text verteilt werden, und *All*, wobei eine Grenze nach jedem Segment eingefügt wird.

In einem nächsten Schritt werden Erweiterungen, in denen wir Wortähnlichkeiten mit Hilfe von *WordNet* testen, eingebaut. Das Ziel dieser Erweiterung ist es, die lexikalischen Ketten nicht nur auf Wortgleichheit aufzubauen, sondern auch auf Beziehungen zwischen Wörtern. Im ursprünglichen System wurden lexikalische Ketten aufgebaut, indem identische Wörter verbunden wurden. Beispielsweise werden im ursprünglichen System nur *house* und *houses* verbunden. Werden Wortähnlichkeiten verwendet, werden auch Wörter wie *houses* und *flat* berücksichtigt. Dies wollen wir für die automatische Themenabgrenzung nutzen, um so die vorhandenen Systeme zu verbessern.

7. Automatische Zusammenfassung

Im Verlauf der letzten 50 Jahre wurden verschiedene Methoden für verschiedene Text- und Zusammenfassungsarten entwickelt. Bei den Methoden gibt es zwei Möglichkeiten: Zusammenfassung durch Extraktion und Zusammenfassung durch Abstraktion. Bei der Extraktion werden Elemente des Textes herausgegriffen und wortwörtlich wiedergegeben. Bei der Abstraktion wird die Zusammenfassung aus Elementen, die nicht im Text vorkommen müssen, gebildet.

Bei den Zusammenfassungsarten gibt es drei prinzipielle Möglichkeiten: eine indikative Zusammenfassung, die die wichtigsten Ideen und Gedanken des Textes wiedergibt, allerdings ohne in die Details zu gehen. Eine themenorientierte Zusammenfassung orientiert sich an den Themen in den Texten, kann sich jedoch auch nur auf ein Thema aus dem Text konzentrieren. Generierte Zusammenfassungen geben auch Standpunkte des Autors wieder.

Textarten, die in der Vergangenheit im Hinblick auf Zusammenfassung untersucht wurden, sind Einzeltexte aus den Bereichen Wissenschaft, Nachrichten aber auch Patentschriften. Auch die Zusammenfassung mehrerer Texte aus einem Themenbereich wurde untersucht. In den letzten Jahren kamen auch Zusammenfassungen von gesprochener Sprache dazu, die sich primär an vorgelesenen Nachrichten bzw. Nachrichtendialogen orientierten.

Wir werden im Folgenden nur auf die Methoden der extraktiven Zusammenfassung eingehen. In frühen Arbeiten wurden satzbasierte Merkmale für die Extraktion verwendet. Diese Merkmale dienten dazu, die zusammenfassungsrelevanten Informationen eines Textes zu finden. Diese Merkmale waren unter anderem die Position des Merkmals im Satz, Schlüsselworte, aber auch häufig wiederholte Wörter und Phrasen. Seit einigen Jahren werden maschinelle Lernverfahren genutzt, um diese Informationen zu finden. Dazu wurde vermehrt Textanalyse eingesetzt, um die Schlüsselworte und -passagen zu finden. Zu dieser Analyse gehören Wortbeziehungen und -ähnlichkeiten, aber auch Diskursstruktur und Anaphernanalyse. Für einen ausführlichen Überblick über die bisherigen Ansätze siehe Radev et al. (2002), Mani und Maybury (1999) und Mani (2001).

Die Zusammenfassung von Diskussionen zwischen mehreren Teilnehmern ist ein sehr neues Forschungsgebiet. Auch wird die Zusammenfassung eher einem Protokoll ähneln als einem fortlaufenden Text, wie bei den textbasierten Ansätzen. Hierfür werden Teile der Diskussionen, die sich zwischen zwei Themengrenzen befinden, als ein Text betrachtet. In nächster Zukunft werden wir extraktive Zusammenfassungsmethoden auf diese Textteile anwenden. Im Bereich der Textzusammenfassung wurden häufig *Lead* und *Tail* als *Baselines* verwendet, die aufgrund der besonderen Struktur von Zeitungstexten auch recht gute Ergebnisse erbringen. Dies ist bei unseren Daten nicht der Fall.

Erfolgversprechendere Möglichkeiten der Zusammenfassung basieren auf Lernverfahren. Hierfür müssen die zusammenfassungsrelevanten Passagen manuell annotiert werden, damit diese dann gelernt werden können. Im Anschluss an die Themengrenzenbestimmung werden zwischen zwei Grenzen die zusammenfassungsrelevanten Segmente oder auch größere Teile der Aussagen einzelner Sprecher manuell annotiert. Diese sollen zum einen dazu dienen, erste Zusammenfassungen zu erstellen und ihre Lesbar- und Verständlichkeit zu prüfen, aber auch um mit Hilfe eines maschinellen Lernverfahrens automatisch zusammenfassungsrelevante Elemente zu bestimmen. Die manuellen Elemente werden dann dazu dienen, die automatischen Verfahren zu evaluieren. Eine zweite Variante der Zusammenfassung nutzt die schon erstellten lexikalischen Ketten. Diese liefern Hinweise auf zusammenfassungsrelevante Begriffe und Phrasen, die bei der Behandlung eines Themas immer wieder vorkommen.

Ein Problem der automatischen Zusammenfassung ist die Evaluierung. In den Bereichen, die bisher vorgestellt wurden, gibt es formale Maße für die

Güte der Ergebnisse. In der Evaluierung von Zusammenfassungen fehlt dieses einheitliche Maß. Es gibt zwei prinzipielle Möglichkeiten: intrinsische und extrinsische Auswertung. Die intrinsische Evaluierung bezieht sich allein auf die Abdeckung der Zusammenfassung als solche. Dabei wird häufig verglichen, inwiefern die automatisch gewählten Passagen mit manuell gewählten Passagen übereinstimmen. Bei der extrinsischen Auswertung wird die Güte der Zusammenfassung im Hinblick auf den Informationsgehalt geprüft. Dabei werden häufig Fragebögen erstellt, bei denen Fragen zum Gesamttext zu beantworten sind, wobei verglichen wird, ob die Fragen mit Hilfe der Zusammenfassung genauso gut beantwortbar sind wie mit Kenntnis des vollständigen Texts.

8. Schluss

Die vorangegangenen Kapitel haben gezeigt, dass die vollautomatische Anaphernresolution in gesprochenen Dialogen und die Zusammenfassung derartiger Dialoge sehr schwierige Aufgaben sind. Für die erste Phase des Projektes DIANA-Summ wurden daher gewisse Vereinfachungen vorgenommen, insbesondere die Verwendung manuell erstellter Transkriptionen anstelle von automatischer Spracherkennung. Für den praktischen Einsatz unter realistischen Bedingungen muss das System aber natürlich auch mit vollständig automatisch erkannter Sprache umgehen können. Es ist daher geplant, im zweiten Teil des Projekts einen automatischen Spracherkennung zu integrieren.

Das hier beschriebene System hat die Form einer Kaskade von einzelnen Komponenten. Fehler, die in vorgeordneten Systemkomponenten (z. B. vom *POS-Tagger* oder Lemmatisierer) gemacht werden, setzen sich durch die folgenden Komponenten hindurch fort, verstärken sich möglicherweise und beeinträchtigen damit das Endergebnis. Die Qualität der Anaphernresolution sowie der Zusammenfassung hängt daher stark von der Qualität der Vorverarbeitungsschritte ab. Die Vorverarbeitungskomponenten beruhen größtenteils auf Methoden des überwachten maschinellen Lernens, welche wiederum von annotierten Daten abhängig sind. Die Qualitätssicherung der zugrundeliegenden Daten hat daher einen hohen Stellenwert.

Danksagung

Das Projekt DIANA-Summ wird gefördert von der Deutschen Forschungsgemeinschaft (STR 545/2-1) und von der Klaus Tschira Stiftung.

Literaturverzeichnis

- Beeferman, Doug/Berger, Adam/Lafferty, John (1999): Statistical models for text segmentation. In: Machine Learning, 34, 177–210.
- Brants, Thorsten (2000): TnT – a statistical Part-of-Speech tagger. In: Proceedings of the 6th Conference on Applied Natural Language Processing, Seattle, Wash., 29 April–4 May 2000, 224–231.

- Brill, Eric (1994): Some advances in transformation-based part-of-speech tagging. In: Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, Wash., 1–4 August 1994, 722–727.
- Carletta, Jean (1996): Assessing agreement on classification tasks: The kappa statistic. In: Computational Linguistics, 22(2), 249–254.
- Galley, Michel/McKeown, Kathleen/Fosler-Lussier, Eric/Jing, Hongyan (2003): Discourse segmentation of multi-party conversation. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Sapporo, Japan, 7–12 July 2003, 562–569.
- Janin, Adam/Baron, Don/Edwards, Jane/Ellis, Dan/Gelbart, David/Morgan, Nelson/Peskin, Barbara/Pfau, Thilo/Shriberg, Elizabeth/Stolcke, Andreas/Wooters, Chuck (2003): The ICSI meeting corpus. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Hong Kong 6–10 April 2003, 364–367.
- Mani, Inderjeet/Maybury, Mark T. (Hg.) (1999): Advances in Automatic Text Summarization. Cambridge, MA.
- Mani, Inderjeet (2001): Automatic Summarization. Amsterdam/Philadelphia.
- Marcus, Mitchell P./Santorini, Beatrice/Marcinkiewicz, Mary Ann (1993): Building a large annotated corpus of English: The Penn Treebank. In: Computational Linguistics, 19(2), 313–330.
- Mieskes, Margot/Strube, Michael (2006): Part-of-Speech tagging of transcribed speech. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), Genua, Italy, 22–28 May, 2006, 935–938.
- Mitchell, Tom M. (1997): Machine Learning. New York.
- Müller, Christoph (2006): Automatic detection of nonreferential *It* in spoken multi-party dialog. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), Trento, Italy, 3–7 April, 2006, 49–56.
- Pevzner, Lev/Hearst, Marti (2002): A critique and improvement of an evaluation metric for text segmentation. In: Computational Linguistics, 28 (1), 19–36.
- Quirk, Randolph/Greenbaum, Sidney/Leech, Geoffrey/Svartvik, Jan (1991). A Comprehensive Grammar of the English Language. London, UK.
- Radev, Dragomir R./Hovy, Eduard/McKeown, Kathleen (2002): Introduction to the special issue on text summarization. In: Computational Linguistics, 28(4), 399–408.
- Schmid, Helmut (1997): Probabilistic Part-of-Speech tagging using decision trees. In: Jones, Daniel/Somers, Harold (Hg.): New Methods in Language Processing. London, UK, 154–164.
- Toutanova, Kristina/Klein, Dan/Manning, Christopher D./Singer, Yoram (2003): Feature-rich Part-of-Speech tagging with a cyclic dependency network. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Edmonton, Alberta, Canada, 27 May–1 June, 2003, 252–259.
- Toutanova, Kristina/Manning, Christopher D. (2000): Enriching the knowledge sources used in a maximum entropy Part-of-Speech tagger. In: Proceedings of the Joint SIG-DAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Hong Kong 7–8 October 2000, 63–70.