Multimodaler Mensch-Maschine-Dialog

C. Fügen, P. Gieselmann, H. Holzapfel, T.Schaaf, A. Waibel
Interactive Systems Labs, ILKD
Universität Karlsruhe
Am Fasanengarten 5
76131 Karlsruhe, Germany
{fuegen,petra,hartwig,tschaaf,waibel}@ira.uka.de

Abstrakt

Dieser Beitrag beschäftigt sich mit dem multimodalen Dialog zwischen Mensch und Roboter. Beschrieben werden die Problematiken im Bereich der Spracherkennung und der Dialogverarbeitung, wobei im Besonderen auf das Erkennen und Erlernen neuer Worte in der Spracherkennung und der Integration von Emotionen in die Dialogstrategie eingegangen wird. Ferner beschreiben wir die bei uns zum Einsatz kommenden Systeme: den Spracherkenner JANUS mit dem neuen Decoder IBIS und das Dialogsystem ARIADNE.

1 Einleitung

Die Schnittstelle zwischen Mensch und Maschine gehört zu den zentralen Komponenten eines Roboters. Sie beeinflusst in großem Maße dessen Akzeptanz beim Menschen, welche bei einem Einsatz eines Roboters im häuslichen Bereich von großer Bedeutung ist. Insofern ist es wichtig, dass sie dem Benutzer die Möglichkeit gibt sich wie mit einem Menschen per Sprache, Gestik und sogar per Mimik unterhalten zu können. Betrachtet man die heutige Forschungslandschaft in diesem Bereich, so sind wir noch weit von dieser Vision entfernt.

In diesem Beitrag beschäftigen wir uns mit dem Bereich der sprachlichen Kommunikation zwischen Mensch und Roboter, d.h. der Spracherkennung, dem Sprachverstehen und der Dialogverarbeitung, wobei wir dabei den multimodalen Aspekt, wie z.B. den Umgang mit Gesten und Emotionen im Dialogsystem, nicht aus den Augen lassen wollen.

Die Problematik der sprachlichen Kommunikation besteht hauptsächlich in der Art der Sprache. Spontansprache und unterschiedliche Dialekte sind für Spracherkennung und Dialogmanagement gleichermaßen ein Problem. So müssen die verwendeten Grammatiken dadurch auftretende grammatikalisch fehlerhafte Konstrukte unterstützen. Die Spracherkennung selber hängt noch von zwei anderen Faktoren, dem Sprecher und den akustischen Bedingungen, ab. Zur Verminderung des Einflusses dieser beiden Faktoren auf die Güte der Spracherkennung werden Adaptionsverfahren angewandt. Von der Güte der gesamten Sprachverarbeitung hängt wiederum die Aufmerksamkeitssteuerung des Roboters ab, wobei zur Verbesserung dessen Robustheit meist auch visuelle Information herangezogen wird.

Da es nahezu unmöglich ist eine bezüglich Wort- und Konzeptumfang vollständige Mensch-Maschine-Schnittstelle zu definieren, muss es einerseits dem Benutzer erlaubt werden neue Worte und Konzepte hinzuzufügen, aber andererseits sollte der Roboter auch in der Lage sein diese selbständig zu erlernen, evtl. auch mit Unterstützung eines zusätzlichen Klärungsdialoges. Die Größe dieses Problems wird in großem Maße auch von der Art der Benutzerführung, d.h. mit Systeminitiative oder mit gemeinsamer Initiative, beeinflusst. Allgemein spielt die Reaktion des Roboters auf Anfragen des Benutzers eine große Rolle bei der Akzeptanzfrage. So lässt sich diese z.B. auch durch Einbindung von Emotionen des Benutzers in die Dialogstrategie und durch emotionale Regungen (sprachliche oder mimische) des Systems verbessern.

Dieser Beitrag gliedert sich wie folgt: In Kapitel 2 gehen wir auf die beschriebene Problematik bei der Spracherkennung ein und widmen uns hauptsächlich dem Aspekt des Erlernens von neuen Worten. In Kapitel 3 beschreiben wir zunächst das bei uns zum Einsatz kommende Dialogsystems, ARIADNE, und den zur Wissensspezifikation verwendeten Formalismus, gehen danach auf die Behandlung multimodaler Eingabeströme ein und untersuchen zum Schluss inwiefern sich Emotionen als Parameter in die Dialogstrategie einbinden lassen. Kapitel 4 schließt den Beitrag mit einer kurzen Zusammenfassung und einem Ausblick auf weitergehende Arbeiten ab.

2 Spracherkennung

Dieser Abschnitt beschäftigt sich hauptsächlich mit der Problematik der Spracherkennung bei der Kommunikation zwischen Mensch und Roboter.

2.1 JANUS und IBIS

Der bei uns zur Verwendung kommende Spracherkenner ist Teil des an der Universität Karlsruhe entwickelten JANUS Recognition Toolkits (JRTk) [3]. Wir verwenden weiterhin den zu JRTk gehörigen, neu entwickelten Single-Pass-Decoder IBIS, der gegenüber der alten Drei-Pass-Suche den Vorteil hat, dass er weniger Speicher benötigt

und zugleich auch schneller ist. Weiterhin erlaubt uns IBIS neben der Verwendung von statistischen n-gram Sprachmodellen (LM) auch entlang kontextfreier Grammatiken (CFG) zu decodieren [4].

2.1.1 Kontextfreie Grammatiken in IBIS

Bei herkömmlichen Spracherkennern, die mit kontextfreien Grammatiken arbeiten, werden aus den Grammatiken endliche Zustandsnetzwerke kompiliert, welche jedoch den Nachteil haben, aufgrund der Menge von Regeln, Knoten und Übergängen mit der Größe der Grammatik überproportional zu wachsen. Jedoch besitzen sie auch den Vorteil, dass die Menge der nachfolgenden Terminalsymbole für einen Knoten direkt bekannt ist und nicht erst während des Decodings ermittelt werden muss. Um beide Vorteile nutzen zu können und es außerdem auch noch zu erlauben, neue Regeln ohne Neukompilation des Netzwerkes hinzufügen zu können, entschieden wir uns bei der Implementierung von IBIS für einen Mittelweg. Wir konstruieren die endlichen Zustandsnetzwerke nur für jede Regel, wobei die Netzwerke unter sich durch die darin enthaltenen Nichtterminalsymbole verbunden sind. Zusätzlich helfen uns intelligente Cachingstrategien, den durch die fehlende Kompaktifizierung entstandenen Overhead zu reduzieren.

Der Vorteil bei der Verwendung von Grammatiken in IBIS ist, dass das Parsing schon während des Decodings erfolgt und somit ein zusätzlicher Parser überflüssig wird. Ferner wird die Möglichkeit unterstützt einzelne Regeln oder sogar ganze Domänen zur Laufzeit vom Decoding auszuschließen oder über einen Gewichtungsfaktor zu bevorteilen. Dadurch haben wir die Möglichkeit von außen den Spracherkennungsprozess auf Teildomänen einzuschränken und erreichen somit in dialogbasierten Anwendungen eine höhere Robustheit.

Grammatiken besitzen gegenüber den statistischen n-gram Sprachmodellen den Nachteil der unzulänglichen Modellierung spontaner nichtsprachlicher sowie sonstiger Geräusche. Dies liegt daran, dass die Position dieser nicht eindeutig vorbestimmt werden kann und die Grammatik es als solche nicht erlaubt, potentiell zwischen allen Worten Geräusche einzufügen. Um diesen Nachteil zu beseitigen modellieren wir Geräusche als sogenannte Filler-Words, die quasi an der Grammatik vorbei in das Decoding integriert werden.

2.2 Sprecher- und sprachliche Merkmale

Die Sprache ist von Mensch zu Mensch aufgrund vieler verschiedener Merkmale sehr unterschiedlich. Zu unterscheiden sind zum einen Sprechermerkmale, wie das Geschlecht, die Stimmlage und die Sprechgeschwindigkeit und zum anderen sprachliche Merkmale, wie die Spontaneität und der Dialekt. Allen Merkmalen gleich ist der negative Einfluss auf die Erkennungsgenauigkeit.

	LM	CFG
WA	76.12%	76.26%
Korrekte Sätze	40.57%	51.23%
RTF on PIII, 1 GHz	0.20	0.16
Speicherbedarf	35 MB	35 MB
Vokabulargröße	2035	2035

Tabelle 1. Vergleich zwischen ein 3-gram LM und einer CFG auf ~250 Sätzen aus der LingWear-Domäne (spontane Anfragen, engl.).

Dem Einfluss dieser Merkmale kann man auf unterschiedlicher Art und Weise entgegenwirken. Man unterscheidet grundsätzlich zwischen modellbasierten und signalbasierten Verfahren, wobei auch die Kombination beider möglich ist.

Die Auswirkungen unterschiedlich langer Vokaltrakte je nach Geschlecht und Stimmlage kompensiert man üblicherweise mit der Vokaltrakt-Längennormierung. Zur Kompensierung spontaner und dialektaler sprachlicher Effekte sowie der Sprechgeschwindigkeit bedient man sich meist modellbasierter Verfahren, wobei hierbei auch die Wahl der Trainingsdaten und das verwendete Sprachmodell eine große Rolle spielen. Hierzu entwickelten wir ein Verfahren, das die in den Trainingsdaten vorhandenen Modalitäten, wie z.B. Sprecher, Geschlecht, und Dialekt besser nutzt, indem rein datengetrieben nicht nur kontextabhängige, sondern auch modalitätenabhängige akustische Modelle trainiert werden. Der Vorteil hierbei ist, dass der Ballungsalgorithmus selbstständig entscheidet, ob die Aufteilung eines Modells entlang einer solchen Modalität gewinnbringend ist, oder nicht. Wir erreichten dadurch Fehlerratenreduktionen von ungefähr 10% [2].

2.2.1 Spontansprache und Grammatiken

In Domänen, wie z.B. bei der Mensch-Roboter-Kommunikation, verwendet man, aufgrund der schon oben genannten Vorteile gegenüber den sonst üblichen statistischen n-gram Sprachmodellen, kontextfreie Grammatiken.

Wir verwenden semantische anstatt syntaktische Grammatiken, da sie robuster gegenüber grammatikalisch fehlerhafte Satzkonstruktionen sind, wie sie gerade in spontaner Sprache auftreten [5]. Jedoch haben sie auch den Nachteil, dass sie üblicherweise schwer auf neue Domänen übertragbar sind, weshalb wir sie deshalb in kleinere Module aufteilen. Jede Teilgrammatik deckt dann alle Konzepte bezüglich einer Teildomäne ab, wobei eine weitere große Grammatik für domänenübergreifende Konzepte verantwortlich ist. Wir erlauben außerdem die Zuweisung von Domänenbezeichnern zu einzelnen Grammatiken, wodurch es uns erlaubt ist, einfach zwischen den Domänen hin- und herzuwechseln, ohne dabei den Spracherkenner neu zu starten.

In Tabelle 1 sind beispielhaft die Ergebnisse von Untersuchungen in der Domäne von LingWear [1], d.h. spon-

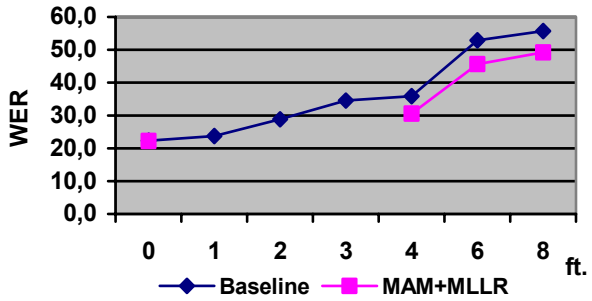

Abbildung 1. Ergebnisse von 9 Sprechern auf distanten Sprachdaten (gelesene Nachrichten, engl.).

tansprachlicher Anfragen für einen tragbaren linguistischen Assistenten für Touristen, aufgeführt. Hieran erkennt man, dass die Grammatik bezüglich der vollständig korrekten Sätze und der Erkennungsgeschwindigkeit (RTF) um etwa 20% besser ist als ein statistisches Sprachmodell.

2.3 Akustische Bedingungen

Gerade ein Roboter im häuslichen Bereich ist vielen unterschiedlichen Umgebungsgeräuschen ausgesetzt. Zusätzlich dazu erzeugen seine Motoren und Gelenke weitere Geräusche, die alle von auf dem Roboter befestigten Mikrophonen aufgenommen werden. Ein weiterer Einflussfaktor auf die Spracherkennung ist die Entfernung zwischen Benutzer und Mikrophon, da dadurch zum einen der Signal-Rausch-Abstand der Aufnahme kleiner wird, und zum anderen Reflexionen des Sprachsignals, hervorgerufen durch die Raumakustik, vermehrt auftreten.

Zur Verminderung der Einflüsse von Nebengeräuschen in Autos und auch zur Kompensation von Kanal- und Sprechereinflüsse entwickelte Martin Westphal im Rahmen seiner Dissertation an der Universität Karlsruhe die sogenannte modellbasierte akustische Transformation, ein Verfahren das Kompensations- mit Adaptionstechniken verbindet [17]. Dieses Verfahren setzen wir auch erfolgreich zur Verbesserung der Spracherkennung mit distanten Mikrophonen ein, wie aus Abbildung 1 ersichtlich ist.

Wichtig ist jedoch in jedem Fall, dass aufgrund der wechselnden akustischen Gegebenheiten adaptive Verfahren zum Einsatz kommen.

2.4 Erkennen und Erlernen neuer Wörter

Während Menschen innerhalb einer Unterhaltung oftmals mühelos ihnen unbekannte Wörter aus dem Kontext des Gesprächs deuten können oder im Zweifelsfalle nachfragen, bereitet diese Eigenschaft Maschinen noch die größten Schwierigkeiten. Gerade bei flexions- und kompositareichen Sprachen, wie z.B. dem Deutschen, ist der zur Verfügung stehende Wortschatz nahezu unendlich, weshalb es für Maschinen bislang unmöglich ist mit einen solch großem Wortschatz umzugehen.

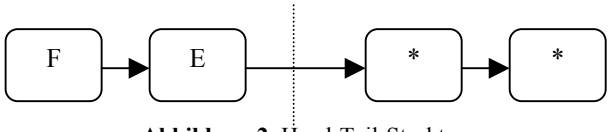

Abbildung 2. Head-Tail-Struktur.

Dennoch sollte ein Roboter in der Lagen sein ihm unbekannte (OOV = out of vocabulary) Wörter zu identifizieren und sie – gegebenenfalls mit Hilfe von Klärungsfragen – seinem Wortschatz und Sprachmodell sinnvoll hinzuzufügen. Dieser Vorgang vollzieht sich in mehreren Schritten. Zunächst müssen unbekannte Wörter aus einer Anfrage als solche identifiziert werden. Danach muss dieses Wort verschriftet und dem Wortschatz hinzugefügt werden und zum Schluss muss die Semantik des Wortes erfasst werden [16][15].

2.4.1 Identifizierung unbekannter Wörter

Treten in einem Satz dem Spracherkenner unbekannte Wörter auf, so werden durch Einfügen, Entfernen und Ersetzen auf Wortebene Erkennungsfehler gemacht. Dies bedeutet jedoch nicht automatisch, dass der Satz als solches keinen Sinn mehr ergibt, was vor allem bei der Verwendungen von Grammatiken als Sprachmodelle häufig auftreten kann. Das am häufigsten angewandte Verfahren, um dieses Problem in den Griff zu bekommen ist, den Spracherkenner so zu modifizieren, dass er anstelle normal trainierter Wörter sogenannte generische Wörter verwendet. Die Problematik hierbei ist jedoch die Konstruktion solcher Wörter, da sie selbst bei spontanen Anfragen unter widrigen akustischen Bedingungen keine richtig erkannten Wörter verdrängen sollten.

Die akustischen Einheiten für generische Wörter werden aus generischen Phonemen gebildet, die aus dem ganzen oder zumindest aus einem Teil des Phoneminventars bestehen. Um der obigen Problematik entgegenzuwirken, verwenden wir zur Bildung generischer Wörter nicht nur generische Phoneme sondern bedienen uns der in Abbildung 2 gezeigten sogenannten Head-Tail-Strukturen. Diese bestehen aus gemischten Ketten normaler und generischer Phoneme. Wie in Tabelle 2 ersichtlich, zeigen unsere Experimente, dass wir hiermit signifikante Verbesserungen für gelesene und spontane Sprache erreichen. Hierbei ist WCE, die Fehlerrate nach Ersetzung unbekannter Wörter in der Referenz und generalisierter Wörter in der Hypothese durch <UNK>. REC und PRC stehen für recall

	WCE	REC	PRC	WCE	REC	PRC
	gelesen			spontan		
BASE	38.9%			22.6%		
GW780	21.1%	59%	100%	22.2%	57%	77%
CHEAT	0.4%	97%	100%	21.9%	74%	100%

Tabelle 2. Ergebnisse auf gelesenen und spontanen Daten (deutsch).

und precision und geben zum einen an wie viele der in den Daten vorkommenden unbekannten Wörter erkannt wurden (REC) und zum anderen ob alle erkannten unbekannten Wörter auch korrekt sind (PRC).

2.4.2 Verschriftung unbekannter Wörter

Damit der Spracherkenner mit den neuen Wörtern umgehen kann, ist es nötig sie zu verschriften. Hierzu bieten sich mehrere Möglichkeiten in Abhängigkeit des schon vorhandenen Wissens über die unbekannten Wörter an.

Üblicherweise geht man, falls die Schreibweise des Wortes bekannt ist wie folgt vor. Zunächst versucht man das Wort in einem großen Hintergrundlexikon zu finden. Ist das nicht möglich so lässt sich mit Hilfe von regelbasierten Graphem-nach-Phonem Konvertierungsverfahren, wie sie z.B. in Sprachsynthesesystemen zum Einsatz kommen, eine Verschriftung generieren. Die Robustheit dieser Verfahren lässt sich noch zusätzlich durch Vergleich mit einer Phonemerkennung auf dem entsprechenden Sprachsegment erhöhen.

Bei unbekannter Schreibweise ist es nur möglich die Verschriftung durch Phonemerkennung zu ermitteln. Ist das zu unsicher kann der Roboter einen Klärungsdialog initiieren, indem er den Benutzer bittet das fragliche Wort zu buchstabieren. Bei der Phonemerkennung erreichen wir mit unseren Ansätzen zur Zeit eine Fehlerrate von 35%.

2.4.3 Erfassung der Semantik unbekannter Wörter

Der Vorteil bei der Verwendung der Head-Tail-Strukturen ist die Möglichkeit diese als Wörter in das Sprachmodell mit aufzunehmen, sodass es mitunter sogar möglich ist, das dieselbe Head-Tail-Struktur in verschiedenen semantischen Klassen des Sprachmodells vorkommt. Wird eine solche Head-Tail-Struktur erkannt, so lässt sich aus der Klassenzuordnung auch auf ihre Semantik schließen. Bei der Verwendung von Grammatiken als Sprachmodellen ergibt sich noch der zusätzliche Vorteil, dass durch den Aufbau der Grammatik gleichzeitig mehrere Kontexte mit diesem Wort ergänzt werden können. In statistischen n-gram basierten Sprachmodellen kann meist nur eine Klasse um dieses Wort ergänzt werden. Falls die Schreibweise des Wortes für den Roboter nicht von Interesse ist, kann man also durch die Klassenzuordnung der Head-Tail-Strukturen direkt auf die Semantik des Wortes schließen.

Die Robustheit und Generalisierbarkeit lässt sich hierbei noch erhöhen, falls die Schreibweise und der momentane Kontext in dem das unbekannte Wort aufgetreten ist, bekannt ist. So kann man z.B. in anderen externem Datenmaterial nach ähnlichen Kontexten suchen und diese bei der Integration des unbekannten Wortes berücksichtigen.

3 Dialogverarbeitung

Damit der Benutzer mit dem Roboter kommunizieren kann, muss er „verstehen“, was der Mensch sagt, um darauf angemessen reagieren können. Die Dialogverarbeitung analysiert die Äußerungen des Benutzers, baut daraus eine semantische Repräsentation auf und übermittelt entweder die auszuführende Anweisung des Benutzer an die nachfolgenden Module, oder fragt bei unterspezifizierten, oder unklar ausgedrückten Anweisungen zurück.

Mögliche Dialogziele in diesem Roboterkontext wären z.B. das Nehmen oder Hinlegen eines Gegenstandes oder auch so komplexe Handlungen wie beispielsweise das Backen eines Pfannkuchens.

Wichtig ist, dass der Benutzer nicht an eine bestimmte Form gebunden ist, um etwas auszudrücken, sondern frei mit dem Roboter reden kann. Es ist also z.B. egal, ob der Benutzer sagt „Gib’ mir mal den roten Becher.“ oder „Ich hätte gern den roten Becher.“ oder „Ich brauche den roten Becher jetzt“. Der Roboter wird in allen drei Fällen aus dem Gesagten erkennen, dass er dem Benutzer den roten Becher bringen soll. D.h., das Ziel „etwas bringen“ wird erkannt sowie auch der Gegenstand, der gebracht werden soll. Was hierbei noch fehlt, ist die Angabe des Ortes, an dem sich der rote Becher befindet, wobei dies entweder visuell oder per Klärungsdialog erfolgen kann.

3.1 Der Dialogmanager ARIADNE

Für das Dialogmanagement wird der an der Carnegie Mellon Universität in Pittsburgh von Matthias Denecke entwickelte sprach- und aufgabenunabhängige Dialogmanager ARIADNE verwendet [9]. Dieser Dialogmanager bietet die oben erwähnten Möglichkeiten der freien Eingabe für den Benutzer. Verwendet werden typisierte Merkmalsstrukturen, die es dem Benutzer erlauben seine Anfragen frei zu formulieren, anstatt wie bei rahmenbasierten Dialogmanagern zu jedem Zeitpunkt genau eine Anfrage zu verfassen.

Darüber hinaus eignet sich ARIADNE besonders gut zum „Rapid Prototyping“ [10], da nur die domänen- und sprachspezifische Komponenten angepasst werden müssen und dabei auf z.B. schon vorhandene generelle Konzepte zurückgegriffen werden kann. Der zugrundeliegende Formalismus mit dem dies erreicht werden kann sind vektorisierte kontextfreie Grammatiken mit ihren Vererbungsmechanismen, die im folgenden noch erläutert werden sollen. Darüber hinaus werden von dem System generelle Ein- und Ausgabemechanismen und Möglichkeiten zur Auswertung des Dialog- und allgemein des Diskurszustandes zur Verfügung gestellt.

ARIADNE arbeitet ferner mit multidimensionalen typisierten Merkmalsstrukturen [11]; auf diese Weise ist es möglich, nicht nur die semantischen Informationen an den Knoten aufzuführen, sondern auch Informationen der Informationsquellen, wie z.B. Konfidenzmaße für die Spracherkennung, generelle Informationen zu den verschiedenen Eingabekanälen, die Anzahl der Versuche, eine bestimmte Information vom Benutzer zu bekommen, usw. Auf diese Weise ist es z.B. möglich, nach einzelnen Worten mit schlechter Konfidenz gezielt nachzufragen

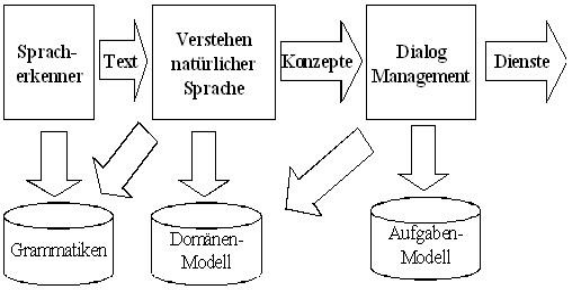

Abbildung 3. Der Dialogmanager und seine Resourcen.

3.1.1 Resourcen des Dialogmanagers

Der Dialogmanager arbeitet mit verschiedenen aufgaben- und domänenabhängigen Wissensquellen. Dazu gehören eine Ontologie, eine Spezifikation der Dialogziele, Datenbankregeln, eine Grammatik und Generierungsschablonen (siehe Abbildung 3). Die Dialogstrategie entscheidet dann wie die vorhandenen Informationen interpretiert und integriert werden.

3.1.2 Grammatik und Domänenmodell

Mit Hilfe der Grammatik wird die Eingabe geparst. Dabei handelt es sich um eine semantische Grammatik, wobei die Nichtterminalsymbole sowohl syntaktische als auch semantische Informationen kodieren. Das Problem dabei ist, dass die semantischen Informationen natürlich domänen- und aufgabenabhängig sind, während die syntaktischen Informationen größtenteils unabhängig von dem einzelnen zu erstellenden Dialogsystem sind. Generelle domänenabhängige Informationen und Vorgehensweisen können dem „Rapid Prototyping“ folgend wiederverwendet werden. Andere Dialogmanager erlauben das aufgrund der fehlenden Trennung von Semantik und Syntax nicht. Ermöglicht wird dies durch sogenannte vektorisierte kontextfreie Grammatiken, bei denen Nichtterminalsymbole aus n-dimensionalen Vektoren über partiell geordneten Elemente bestehen [12].

Durch die Separierung von semantischer und syntaktischer Information ist es möglich syntaktischen Informationen wiederzuverwenden, wie beispielsweise auch der syntaktische Aufbau von komplexeren Nominalphrasen, wie „der rote Becher“. Solche Informationen werden daher in einem generellen domänenunabhängigen Teil der Grammatik festgelegt, während die wirkliche semantische Instantiierung erst in dem konkret zu erstellenden Grammatikteil definiert wird.

Außerdem ist es dadurch auch möglich, die eigentliche Grammatik mit dem Domänenmodell zu kombinieren. Im Domänenmodell wird festgelegt, welche Konzepte das System kennt und wie sie verknüpft werden können. Das Domänenmodell ist dabei als Ontologie konzipiert, sodass Objekte, Aktionen und Eigenschaften, die dort spezifiziert

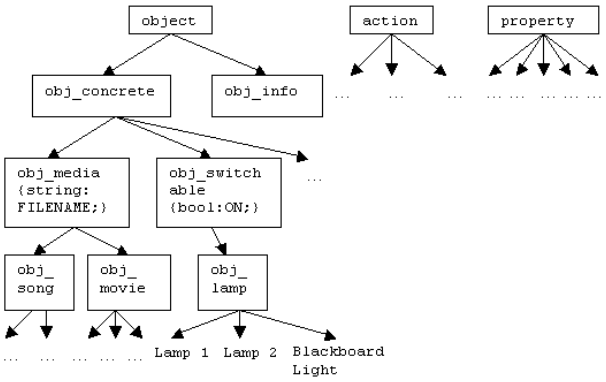

Abbildung 4. Ein Teil einer Ontologie. Die generischen Konzepte werden hier mit anwendungsspezifischen Informationen verbunden.

werden, jeweils von anderen Objekten erben können. Auf diese Weise kann auch auf eine allgemeinere Ontologie (siehe Abbildung 4) zurückzugegriffen werden, die beispielsweise Konzepte wie verschiedene Sprechakte und generelle Ziele, Objekte und Eigenschaften, aus denen dann die spezifischen Objekte, Aktionen und Eigenschaften im domänenabhängigen Teil entwickelt werden können, enthält.

Diese vom Dialogmanager verwendete Grammatik kann auch wieder in eine nicht-vektorisierte Form konvertiert werden, wodurch Spracherkenner und Dialogmanager auf eine gemeinsame linguistische Wissensbasis zugreifen können.

3.1.3 Aufgabenmodell

Das Aufgabenmodell spezifiziert die Dialogziele, die den Diensten, die vom System ausgeführt werden können, entsprechen. Ein Dialogziel kann dabei gesehen werden als die Beschreibung eines Formulars, das mit Hilfe des Dialogs zwischen Mensch und Maschine ausgefüllt wird [9].

Ein Dialogziel ist somit definiert über die Informationen, die vom Benutzer im Diskurs gegeben werden, und setzt sich aus den Objekten, Aktionen und Eigenschaften, die in der Ontologie definiert sind, zusammen. Dabei spielt es keine Rolle, welche Informationen der Benutzer zuerst gibt und welche später. Dadurch ist es z.B. auch möglich zunächst etwas über den Becher zu sagen und erst dann zu erläutern was damit gemacht werden soll. Somit stellen die Dialogziele die Verbindung zwischen den Konzepten des Domänenmodells und den Diensten, die der Dialogmanager ausführen kann, dar.

Ist ein Dialogziel erkannt, sucht der Dialogmanager im Diskurs nach den entsprechenden Variablen, wie Objekte, Eigenschaften und Aktionen. Sollte die Merkmalsstruktur noch unterspezifiziert sein, initiiert der Dialogmanager einen Klärungsdialog.

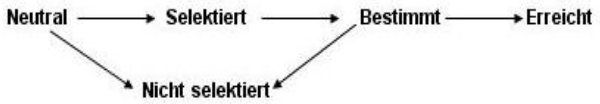

Abbildung 5. Dialogziel-Zustände und ihre Übergänge.

3.1.4 Generierungsschablonen

Klärungsfragen oder andere Informationen werden vom System in natürlicher Sprache formuliert, wobei hierzu sogenannte Generierungsschablonen verwendet werden. In diesen Schablonen wird abhängig vom aktuellen internen Zustand oder einem eingetretenen Ereignis definiert, was der Dialogmanager ausgeben soll. Der Dialogzustand wird dabei definiert über die vorhandene Information in den Dialogzielen (Abbildung 5) [9]. Der Dialogzustand ist zunächst *neutral* am Beginn des Dialogs. Werden dann mehrere Dialogziele ausgewählt, ist er *selektiert*. Ein Dialogziel im selektierten Zustand, wird *bestimmt*, wenn es nur noch als einziges ausgewählt ist. Schließlich ist das Dialogziel *erreicht*, wenn alle Informationen, die im Dialogziel spezifiziert sind, im Diskurs vorhanden sind.

Mit diesen Schablonen wird auf der einen Seite definiert, bei welchem Informationsstand der Dialogmanager wie nachfragt und auf der anderen Seite wird festgelegt, was der Dialogmanager als Antwort vom Benutzer erwartet und wie diese Antwort in das aktuelle Dialogziel integriert werden kann.

Auch bei Robotern ist es wie beim Menschen nötig, dass Informationen auch aus dem Kontextwissen heraus erschlossen werden, sodass keine unnötigen Klärungsfragen generiert werden müssen. So sollte der Roboter wissen, wenn nach einem Becher gefragt wird und nur ein Becher im Raum vorhanden ist, dass dieser Becher damit gemeint ist.

3.1.5 Datenbanken

Die Datenbank enthält Objekte mit ihren Eigenschaften, wie sie z.B. in der Umgebung des Roboters vorkommen, sodass der Dialogmanager in der Lage ist dort z.B. nach den verschiedenen Instanzen des Becher-Objektes nachzuschlagen und somit auch die Information über ihren aktuellen Aufenthaltsort bekommt.

Die Datenbank lässt sich sowohl in Access definieren als auch durch eine Java-Anwendung simulieren. Der Vorteil der Java-Implementierung liegt darin, dass es dort einfacher ist veränderliche Zustände von Objekte zu handhaben. Wenn z.B. eine Lampe bereits angeschaltet ist, und der Benutzer trotzdem den Roboter bittet, die Lampe anzuschalten, wird der Roboter wohl zurückfragen: „Die Lampe ist bereits an. Soll ich sie heller machen?“.

3.1.6 Dialogstrategie

Die Dialogstrategie schließlich definiert, wie die verschiedenen Informationen in einem bestimmten Dialogzustand

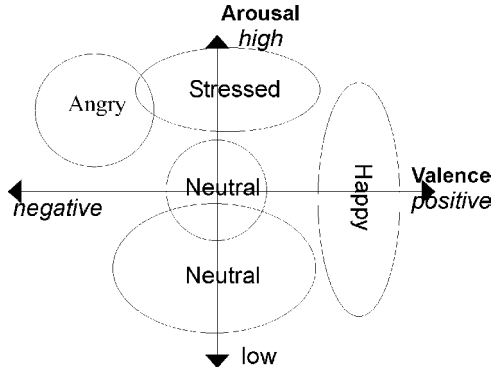

Abbildung 6. Arousal - Valence Ebene.

ausgewertet werden sollen. Hier wird also die generelle Vorgehensweise festgelegt, wie beispielweise mit zu schlechten Konfidenzen des Spracherkenners umgegangen werden soll. Die Dialogstrategie setzt sich dabei aus Interaktionsmustern zusammen, die festlegen, wie Informationen zum Diskurs hinzugefügt bzw. entfernt werden können.

Die Dialogstrategie muss nicht für jede Applikation neu erstellt werden, sondern lässt sich vielfach wiederverwenden. Neue Dialogstrategien sind z.B. zu entwickeln, wenn zusätzliche Modalitäten, wie z.B. Emotionen integriert werden sollen.

3.2 Multimodaler Parser

Da Kommunikation nicht nur aus Sprache besteht, sondern auch aus Gestik, soll ein Roboter auch in der Lage sein, beide Eingabemodalitäten zu verstehen. Diese werden durch einen multimodalen Parser zusammengeführt. Dabei wird über die Zeitpunkte, zu denen Gestik und Sprache auftreten, unifiziert. D.h., tritt beides näherungsweise zum gleichen Zeitpunkt auf, geht man davon aus, dass es zusammen ausgewertet werden soll.

Auch hierbei ist es – ähnlich wie beim Sprachsignal – möglich, die Konfidenzen, mit denen auf einen bestimmten Punkt gedeutet wurde, mitauszuwerten und wenn nötig, nachzufragen, welchen Gegenstand der Benutzer wirklich meinte. Gleichzeitig können sich natürlich die beiden Modalitäten auch ergänzen. D.h., wenn beispielsweise sowohl die Spracherkenner-Hypothese als auch die Hypothese vom Gestenerkenner nur eine geringe Konfidenz aufweisen, beide gleichzeitig aber auf den gleichen Gegenstand referieren, wird die Gesamtkonfidenz dadurch wieder aufgewertet.

Ebenso können durch Fusion von Sprache und Gestik Ambiguitäten aufgelöst werden, wenn der Benutzer z.B. sagt, „Gib mir bitte den Becher“ und gleichzeitig auf einen bestimmten Becher zeigt, ist für das System klar, welcher Becher gemeint ist. Außerdem bietet es sich natürlich auch an, wie in [7] dargestellt über die Semantik von Gestik und Sprache zu unifizieren.

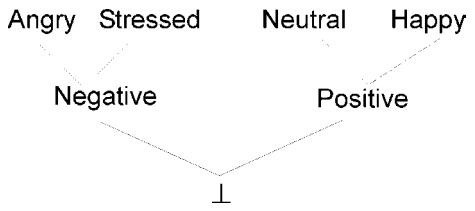

Abbildung 7. Diskretisierte Darstellung.

3.3 Emotionen als Dialogstrategieparameter

Durch die Integration von Emotionen in ein Dialogsystem wird es möglich neue Strategien zu definieren, die besser auf den Zustand des Benutzers angepasst sind. Während es auch Arbeiten gibt, die sich mit System-Emotionen befassen, betrachten wir in dieser Arbeit nur Emotionen des Benutzers. System-Emotion werden benutzt um dem Benutzer Informationen auf andere Art und Weise zu vermitteln. Unser Ansatz, Benutzer-Emotionen zu modellieren, ermöglicht es dem System, abhängig vom Zustand des Benutzers andere Strategien auszuwählen. Diese Strategien entscheiden welche Informationen erfragt werden und wie diese Information im gegebenen Kontext interpretiert wird.

3.3.1 Eigenschaften von Emotionen

Um Emotionen im Dialogsystem zu verwenden, ist es nötig, eine möglichst robuste Erkennung der Emotionen bereitzustellen. Die Merkmale, die dazu verwendet werden, müssen mit verfügbaren Sensoren gemessen werden, und das Resultat des Erkenners muss semantisch sinnvolle Information für die Anwendung liefern.

In der Literatur finden sich verschiedene Werke, die Eigenschaften von Emotionen und mögliche Algorithmen diese zu erkennen beschreiben, z.B. [14]. Es gibt hauptsächlich zwei unterschiedliche Ausprägungen von Emotionen, die beim Menschen auftreten. Diese sind physisch und kognitiv. Dementsprechend gibt es auch generell zwei Ansätze um Emotionen zu erkennen.

Um Emotionen nach einem kognitiven Modell zu schließen ist ein umfangreiches Weltmodell nötig, da es Wünsche, Ziele und Ängste des Benutzers in betracht ziehen muss [13]. Situationen des alltäglichen Lebens sind allerdings zu komplex, um sie korrekt zu modellieren. Damit erfüllt das Modell nicht die Voraussetzungen, um gute Rückschlüsse auf die Emotionen von Menschen machen zu können.

Ein weiterer Ansatz misst die physischen Veränderungen des Menschen, welche über Sprache, Mimik, Herzschlag und viele weitere zu erkennen sind [14]. Dieser Ansatz ist der von uns favorisierte, da er eine stärker Domänen-unabhängige Methodik zur Verfügung stellt. Standardmäßig werden Emotionen in einem 2-dim. Koordinatensystem dargestellt, mit den Achsen arousal und valence [8], siehe Abbildung 6. Die vom Dialogsystem verlangte diskrete Repräsentation ist in Abbildung 7 dargestellt.

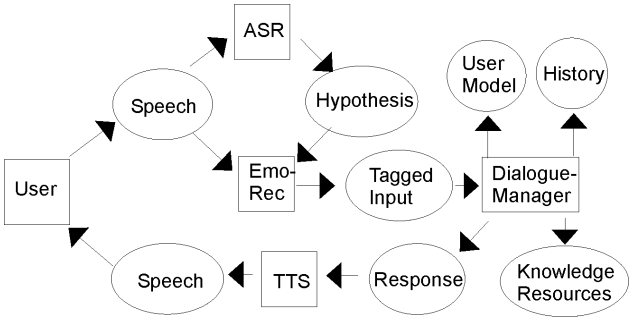

Abbildung 8. Datenfluss und beteiligte Komponenten.

Allerdings abstrahiert der Entwurf des Dialogsystems und die Einbindung der Emotionen von dem tatsächlichen Emotionsmodell [6]. Damit wird es möglich sein, in unserem System einen erweiterten Emotionserkenner aufzunehmen, der neben der Klassifiktaion auf Signalebene auch kognitive Hinweise verwendet, wenn es sich herausstellt, dass damit die Akkuratheit des Modells verbessert werden kann.

3.3.2 Architekturelle Einbindung in das Dialogsystem

Abbildung 8 zeigt die beteiligten Komponenten des Dialogsystems in einem Datenfluss-Diagramm. Die Eingabe ist in der dargestellten Konfiguration rein sprachbasiert. Die Hypothesen von Spracherkenner und Emotionserkenner werden in eine semantische Repräsentation umgewandelt und an das Dialogsystem weitergeleitet Durch die bereits erwähnten multidimensionalen typisierten Merkmalsstrukturen ist es möglich, die Sprache des Benutzers mit Emotions-Information zu annotieren [6].

3.3.3 Erweiterung der Dialog-Strategie

Der Dialogmanager verfügt, wie bereits erwähnt, über einen Katalog von Interaktionsmustern. Diese werden verwendet um Information dem Diskurs hinzuzufügen oder daraus zu entfernen. Die Applikation beschreibt verschiedene Instanzen der Interaktionsmuster und spezifiziert die Vorbedingungen und die Aktionen für deren Ausführung.

Die Auswahl der Interaktionsmuster und die Vorbedingungen werden auf Basis des abstrakten Dialogzustandes spezifiziert [9]. Der abstrakte Dialogzustand enthält Variablen, die Informationen über die Eingabe und den Dialogverlauf enthalten. Mit der Einführung von neuen Variablen kann der abstrakte Dialogzustand erweitert werden, um auch Emotionswerte abzubilden. Die folgenden drei Variablen werden für Emotionen verwendet:

- UserEmotion: Modelliert den aktuellen emotionalen Zustand des Benutzers.
- SystemEmotion: Repräsentiert die Strategie des Systems, wie es auf die UserEmotion reagiert.
- EmotionState: Modelliert die emotionale Tendenz des Benutzers.

Während die UserEmotion nur den emotionalen Zustand des Benutzers zum Zeitpunkt der Eingabe modelliert, modelliert EmotionState die Tendenz des Benutzers.

Mit dem EmotionState lässt sich z.B. erkennen, ob der Benutzer sich über längere Zeit über das System ärgert. Durch diese Möglichkeit können Strategien definiert werden, die besser auf den Zustand des Benutzers angepasst sind.

3.3.4 Applikationen

Wie zu Anfang des Kapitels notiert, soll das Ziel der Anwendung sein, eine möglichst sinnvolle Reaktion auf die Eingabe des Benutzers zu zeigen, abhängig auch vom emotionalen Zustand des Benutzers. Das bedeutet nicht unbedingt, menschliches Verhalten nachzuahmen. Zum Beispiel ist es unklar, ob ein Roboter oder Computer jemals Wutausbrüche haben sollte. Vielmehr kann eine wütende Reaktion des Benutzers ein Indiz für Fehlverhalten des Roboters sein. Dementsprechend wäre es die Aufgabe des Dialogsystems, die auszuführende Aufgabe durch Rückfragen an den Benutzer zu überprüfen.

4 Zusammenfassung und Ausblick

In diesem Beitrag haben wir einen Überblick über unsere aktuellen Arbeiten im Bereich des multimodalen Mensch-Maschine-Dialogs gegeben. Wir haben die Problematik bei der natürlichsprachlichen Kommunkation von Mensch und Maschine innerhalb der Spracherkennung, des Dialogmanagements und des Sprachverstehens aufgezeigt. Ferner sind wir im speziellen auf die Verwendung und Modellierung von kontextfreien Grammatiken in Dialog und Spracherkennung eingegangen und haben uns der Thematik gewidmet, mit welchen Verfahren unbekannte Worte erkannt und erlernt werden können.

Hierbei ausgelassen haben wir die grundsätzliche Problematik des Lernens von ganzen Grammatiken und ihren Konzepten zuzüglich der benötigten Dialogziele im Dialogmanager. Damit wollen wir uns verstärkt in zukünftigen Arbeiten beschäftigen.

In der Dialogverarbeitung haben wir uns mit Emotionen als Parameter in der Dialogstrategie beschäftigt. Auch dieser Ansatz soll weiterverfolgt werden; bietet er doch die Möglichkeit, die Dialogstrategie auf den Benutzer anzupassen und damit die Akzeptanz des Systems beim Benutzer zu erhöhen.

Der multimodale Aspekt – in diesem Fall die Berücksichtigung visueller Informationen, wie z.B. Gestiken – ist nur angerissen worden und soll auch für zukünftigen Arbeiten erweitert werden.

5 Danksagung

Diese Arbeit ist Teil des Sonderforschungsbereichs (SFB) 588, „Humanoide Roboter – Lernende und kooperierende multimodale Roboter“ an der Universität Karlsruhe. Der SFB wird von der deutschen Forschungsgemeinschaft (DFG) unterstützt.

6 Referenzen

[1] C. Fügen, M. Westphal, M. Schneider, T. Schultz, A. Waibel: *LingWear: A Mobile Tourist Information System*. In Proc. of the Human Language Technology Conference, HLT-2001, San Diego, March 2001.

[2] C. Fügen, I. Rogina*: Integrating Dynamic Speech Modalities into Context Decision Trees*. In Proc. of the International Conference of Acoustics, Speech and Signal Processing, ICASSP-00, Istanbul, Turkey, June 2000.

[3] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal: *The Karlsruhe-Verbmobil Speech Recognition Engine*. In Proc. of the International Conference on Acoustics, Speech and Signal Processing, ICASSP-97, Munich, Germany, 1997.

[4] H. Soltau, F. Metze, C. Fügen and A. Waibel: *A One pass-Decoder based on Polymorphic Linguistic Context Assignment*. In Proc. of the Automatic Speech Recognition and Understanding Workshop, ASRU-2001, Madonna di Campiglio, Trento, Italy, December 2001.

[5] M. Woszczyna, M. Broadhead, D. Gates, M. Gavaldà, A. Lavie, L. Levin, A. Waibel: *A Modular Approach to Spoken Language Translation for Large Domains*. In Proc. of AMTA-1998.

[6] H. Holzapfel, C. Fügen, M. Denecke, A. Waibel: *Integrating Emotional Cues into a Framework for Dialogue Management*. In Proc. of the International Conference for Multimodal Interfaces, ICMI-2002, Pittsburgh, October 2002.

[7] F. Landragin: *The Role of Gesture in Multimodal Referring Actions*. In: Proceedings of the 4th International Conference on Multimodal Interfaces, 2002.

[8] P. J. Lang, M. M. Bradley, B. N. Cuthbert: *Emotion, Attention, and the Startle Reflex*, Psychological Review, Vol. 97, No. 3, 1990, pp. 377-395.

[9] M. Denecke: *Generische Interaktionsmuster für aufgabenorientierte Dialogsysteme,* Dissertation, Karlsruhe, 2002

[10] M. Denecke: *Rapid Prototyping for Spoken Dialogue Systems*. In: Proceedings of the 19th International Conference on Computational Linguistics, Taiwan, 2002.

[11] M. Denecke and J.Yang: *Partial Information in Multimodal Dialogue*. In: Proceedings of the International Conference on Multimodal Interfaces, 2000.

[12] M. Denecke: *Object-oriented Techniques in Grammar and Ontology Specification*. In: Proceedings of the Workshop on Multilingual Speech Communication, 2000

[13] A. Ortony, G. L. Clore, A. Collins: *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge, MA, 1988

[14] R. Picard, *Affective Computing*, The MIT Press, 1997.

[15] I. Rogina, T. Schaaf: *Lecture and Presentation Tracking in an Intelligent Meeting Room*. In Proc. of the International Conference on Multimodal Interfaces, ICMI-02, Pittsburgh, USA, October 2002.

[16] T. Schaaf: *Detection of OOV Words using Generalized Word Models and a Semantic Class Language Model*. In Proc. of the EUROSPEECH-01, Aalborg, Denmark, September 2001.

[17] M. Westphal, A. Waibel: *Model-Combination-Based Acoustic Mapping*. In Proc. of the International Conference of Acoustics, Speech and Signal Processing, ICASSP-01, Salt Lake City, USA, May 2001.