Kommandierung eines Serviceroboters mit natürlicher, gesprochener Sprache

C. Fischer, P. Havel, G. Schmidt

Lehrstuhl für
Steuerungs- und Regelungstechnik
Technische Universität München
D-80290 München

J. Müller, H. Stahl, M. Lang

Lehrstuhl für
Mensch-Maschine-Kommunikation
Technische Universität München
D-80290 München

email: {fischer, havel, gs, mue, sta, lg}@{lsr, mmk}.e-technik.tu-muenchen.de

Kurzfassung. Dieser Beitrag beschreibt die Verwendung natürlicher, gesprochener Sprache zur Kommandierung eines als persönlicher Assistent eingesetzten Serviceroboters. Im ersten Schritt wird dazu aus der gesprochenen Äußerung der Bedeutungsinhalt extrahiert. Bei der anschließenden Generierung der Roboterbefehle werden falsche Anweisungen abgewiesen und fehlende Information durch ein Umgebungsmodell oder über eine Rückfrage an den Bediener ergänzt. Das Verfahren ist Bestandteil der Mensch-Roboter-Schnittstelle des mobilen Serviceroboters ROMAN.

1 Einführung

Mobile Serviceroboter gewinnen zunehmend an Bedeutung. Bereits heute umfaßt ihr Aufgabenbereich die Essensverteilung in Krankenhäusern oder die Reinigung von Flugzeughüllen. Weitere Einsatzgebiete liegen im Dienstleistungsbereich als persönlicher Assistent zur Übernahme einfacher Botenaufgaben und zur Unterstützung älterer oder behinderter Menschen bei manuellen Tätigkeiten, sowie im Bereich des Teleservice und der Telediagnose von industriellen Anlagen. Im Idealfall führt der Roboter seinen Auftrag vollkommen autonom aus. Mit den derzeit zur Verfügung stehenden technischen Mitteln aus dem Bereich der künstlichen Intelligenz ist jedoch noch keine selbstständige Durchführung komplexer Serviceaufgaben in unbekannten Situationen möglich. Einen realistischen Kompromiß auf dem Weg zu erhöhter Autonomie bilden semiautonome Robotersysteme, welche bei Bedarf durch das Wissen und die Entscheidungsfähigkeit des Bedieners unterstützt werden. Dabei kommandiert, überwacht und unterstützt der Bediener den Serviceroboter über eine geeignete Mensch-Roboter-Schnittstelle (huMan-Robot-Interface - MRI), siehe Abbildung 1.

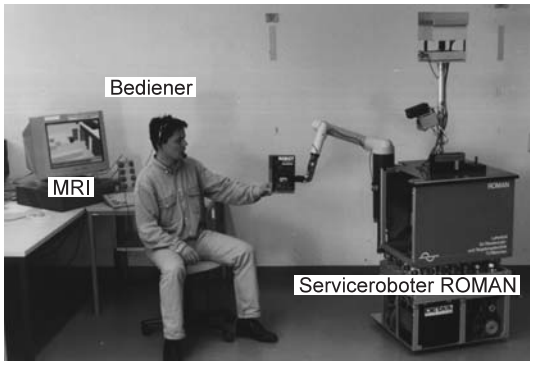

Abb. 1. Serviceroboter ROMAN mit Bediener am MRI

Mit dem Vordringen der Serviceroboter in die unmittelbare Umgebung des Menschen ändern sich im hohen Maße die Anforderungen an die Mensch-Roboter-Kommunikation. Zur Steigerung der Akzeptanz kommt es dabei vor allem darauf an, Menschen auf die gleiche Art und Weise mit dem Roboter kommunizieren zu lassen, wie sie es von einem Mensch-zu-Mensch-Dialog gewohnt sind. Eine unkomplizierte, menschengerechte Kommunikation zeichnet sich durch die folgenden, an die Realisierung gestellten Anforderungen aus: 1. Dialoggeführte natürlichsprachliche und sprecherunabhängige Kommandoeingabe. 2. Visuelle bildschirmgeführte Überwachung und Roboterunterstützung. 3. Haptische Überwachung und Unterstützung bei der mobilen Handhabung. 4. Gesprochener Kommentar bei der Ausführung von Aufgaben.

Der vorliegende Beitrag konzentriert sich auf die sprecherunabhängige Kommandierung eines Serviceroboters mit natürlicher, gesprochener Sprache. In der Literatur sind mehrere Ansätze zur sprachlichen Steuerung eines Roboters bekannt: In [1] wird die Notwendigkeit nach einer einfachen, für den Menschen gewohnten Kommandoeingabe betont. Verwendung findet ein sprecherabhängiger Spracherkenner, mit dem im System vorhandene Makropakete ausgewählt werden können. Die Repräsentation natürlicher Sprache in Bezug auf eine Roboterumgebung bildet den Schwerpunkt der in [2] vorgestellten Arbeit. Dabei wird aus einer textuellen Eingabe direkt ein Roboterbefehl generiert. Eine Unterscheidung der Spracheingabe in Standardeingabe und Telesteuerung nimmt [3] vor. Mit der Interpretation räumlicher Ausdrücke bei der Kommandierung eines Montageroboters beschäftigt sich [4]. Die Übermittlung von Anweisungen bei der Montage wird in [5] durch eine begrenzte Anzahl interner Sensoren gelöst. Dieses Vorgehen bietet die Möglichkeit, fehlende Informationen gezielt nachzuliefern. Trotz vieler Teilergebnisse sind leistungsstarke Gesamtsysteme zur Kommandierung eines Serviceroboters mit sprecherunabhängiger, natürlicher, gesprochener Sprache noch weitgehend unbekannt. Am Beispiel der Einbindung eines bestehenden sprachverstehenden Systems [6] in die Domäne des Serviceroboters ROMAN [7] wird in dem vorliegenden Beitrag die Umsetzung eines akustischen Sprachsignals in Roboteranweisungen aufgezeigt, siehe Abbildung 2.

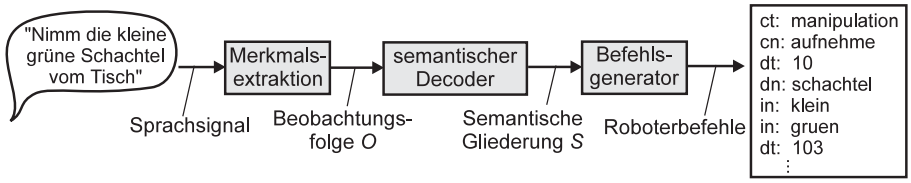

Abb. 2. Natürlichsprachliche Mensch-Roboter-Schnittstelle

Im folgenden Abschnitt wird der Sprachumfang der Roboterdomäne festgelegt und auf die Befehlsschnittstelle zum Serviceroboter eingegangen. Fragen der Decodierung des semantischen Inhalts eines Serviceauftrags behandelt Abschnitt 3 im Zusammenhang mit trainierten, probabilistischen Wissensbasen. Abschnitt 4 beschreibt die nachfolgende Umsetzung der semantischen Gliederung in roboterverständliche Anweisungen. Anhand eines typischen Szenarios aus dem Bereich persönlicher Assistenz wird in Abschnitt 5 die Tragfähigkeit und Robustheit des Gesamtsystems aufgezeigt. Eine Zusammenfassung mit Ausblick schließt den vorliegenden Beitrag.

2 Sprach- und Befehlsumfang der Roboterdomäne

Der Umfang einer Sprache wird durch ihren Wortschatz und die zur Kombination der einzelnen Wörter herangezogene Grammatik definiert. Die Komplexität der deutschen Sprache läßt sich leicht anhand der ca. 300.000 verwendeten Wörter erahnen. Da es neben dem Menschen kein sprachverstehendes System gibt, welches in der Lage ist, einen solch großen Sprachumfang zu verstehen, müssen technische Systeme sich somit auf einen relevanten Teil der Sprache konzentrieren. Dabei ist es sinnvoll, nicht etwa einen repräsentativen Querschnitt durch die menschliche Sprache zu bilden, sondern vielmehr einen Teilbereich (Domäne) möglichst vollständig abzudecken.

2.1 Sprachumfang der Roboterdomäne

Bei der in diesem Beitrag verwendeten Domäne handelt es sich um das Arbeitsfeld eines Serviceroboters zur Ausführung mobiler Handhabungsaufgaben in Innenraumumgebungen. Der Sprachumfang leitet sich dabei von den verschiedenen an den Roboter gestellten und an den anatomischen und planerischen Fähigkeiten orientierten **Serviceaufträgen** ab: 1. Aufnehmen und Abstellen von Objekten, wie z.B. Geschirr, Bücher oder Werkzeug. 2. Bedienen von Einrichtungsgegenständen, wie z.B. Türen, Schränke oder Schalter. 3. Transportaufgaben. 4. Kontinuierliche Bearbeitung auf Boden, Wand oder Mobiliar. Neben der eigentlichen Aktion können vom Bediener eine Fülle von Gegenständen, Räumlichkeiten und symbolischen Positionen in einem Auftrag miteinander kombiniert werden. Dies beinhaltet auch die Verwendung bedeutungstragender Attribute, wie Adjektive, oder sogennanter bedeutungsloser Füllwörter, wie z.B "bitte". Außerdem hat der Bediener die Möglichkeit, den Serviceauftrag durch Verwendung von Synonymen oder unterschiedlicher Detaillierung auf verschiedene Art und Weise zu spezifizieren. So gibt der Bediener mit "Nimm die kleine grüne Schachtel vom Tisch" einen sehr detaillierten Befehl, wohingegen bei "Nimm die grüne Schachtel bitte" von einem Grundwissen über die Position seitens des Roboters ausgegangen wird. Fehlt dieses Grundwissen oder ist der Befehl unvollständig, so muß die fehlende Information durch Rückfrage an den Bediener ergänzt werden.

Abbildung 3 zeigt die Einbettung des Sprachverstehens in die Eingabeseite eines bestehenden MRI. Etwaige Rückfragen an den Bediener erfolgen über die Ausgabeseite genauso wie die für die visuelle Beurteilung durch den Bediener wichtige Darstellung aufbereiteter Sensorinformation. Vom MRI aus lassen sich die vom Planer des Serviceroboters bereitgestellten Fähigkeiten über einen definierten Befehlssatz ansprechen. Durch die Kombination bereits existierender Funktionalitäten zu komplexeren Aufträgen, den sogenannten Makros, können die Funktionalitäten des Gesamtsystems (Serviceroboter und MRI) durch das MRI weiter ausgebaut und somit wachsenden Anforderungen der Domäne angepaßt werden. Es ist auch möglich, die neu definierten Makros in die Wissensbasis der Planungsebene nachzuladen und dadurch den Befehlsumfang des Serviceroboters zu erweitern. In der Planungsebene werden dann durch Kombination von abgelegtem Wissen mit aktueller Sensorinformation Bewegungen generiert. Dabei kann es vorkommen, daß die Fähigkeiten des Planers nicht ausreichen, den

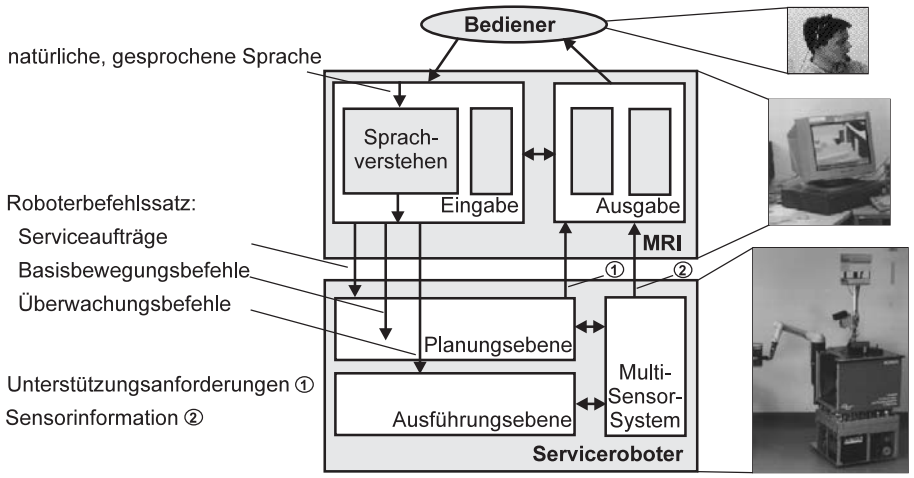

Abb. 3. Ankopplung des Sprachverstehens an einen Serviceroboter

Auftrag erfolgreich auszuführen, und somit eine Unterstützung durch den Bediener erfordern. Ferner sind Situationen denkbar, in denen der Bediener aus Sicherheitsgründen sofort eingreifen muß. Aus diesen Anforderungen heraus lassen sich neben den **Serviceaufträgen** zwei weitere Kategorien von Befehlen, die durch den Sprachumfang beschrieben werden müssen, unterscheiden: Roboterspezifische **Basisbewegungsbefehle** zur Unterstützung, mit denen die Komponenten des Roboters (wendiges Fahrzeug, Effektor, Kamera) direkt angesprochen werden: *"Bewege Dich etwas nach links", "Öffne den Greifer", "Schwenke die Kamera nach unten", ...* **Überwachungsbefehle** zur sofortigen Beeinflussung des Bewegungsablaufs: *"Stop", "Mach' weiter", "Brich ab", ...*

Aus den drei Befehlskategorien resultiert der gesamte Sprachumfang, mit dem eine natürliche, d.h. dem Menschen vertraute, Kommunikation zwischen Mensch und Roboter ermöglicht wird. Der derzeitig verwendete Wortschatz beträgt 409 Wörter, mit dem ungefähr 30 verschiedene Einzelaktionen unterstützt werden.

2.2 Erweiterbare Roboterbefehlsstruktur

Zur Repräsentation der verschiedenen Befehle wird eine für den Planer verständliche Struktur gewählt, mit der sich ohne Änderung neue Befehle implementieren lassen. Jeder Befehl besteht aus einem Header mit Befehlstyp *ct* und Befehlsname *cn*, gefolgt von einer aktionsspezifischen Anzahl von Datenblöcken. Mit Hilfe der Datenblöcke werden die Befehle näher beschrieben. So läßt sich je nach Aktion ein Objekt, die dazugehörige Objektposition oder eine Zielposition unterscheiden. Andere Befehle benötigen wiederum nur eine Mengenangabe oder Zusatzinformation. Um bei komplexeren Befehlen, den Makros, auch unterschiedliche Zielpositionen unterscheiden zu können, ist die Reihenfolge genauso wie die Anzahl der Datenblöcke durch den Befehl vorgegeben. Jeder dieser Datenblöcke besteht aus einem Datentyp *dt*, einem Datennamen *dn* und weiteren Datenelementen. Mögliche Datenelemente sind die Datenmenge *da*, ein Positionsvektor *po* und Zusatzinformationen *in*. Das folgende Beispiel repräsentiert die Äußerung *"Nimm die kleine grüne Schachtel vom Tisch"*. Die nebenstehende Tabelle listet alle zur Verfügung stehenden Datentypen *dt* auf:

```
[
  [ct:mobile_manipulation][cn:aufnehme]
  [
    [dt:10][dn:schachtel][da:1][in:klein][in:gruen];
    [dt:103][dn:tisch][po:3540,2000,800,0,0,0]
  ]
]
```

dt	name
10	object
101	relative location
102	absolute location
103	object referenced location
20	goal
201	relative location
202	absolute angle value

dt	name
203	absolute location
204	object referenced location
30	relative amount
31	identifier
32	force amount
33	length amount
40	information

3 Semantische Decodierung

Spracherkenner zur Decodierung der Wortkette einer gesprochenen Äußerung sind nicht neu. Selbst die sprecherunabhängige Verarbeitung fließender Sprache ist bei eingeschränktem Vokabular möglich. Weitestgehend ungelöst ist jedoch die Einbeziehung semantischer oder gar pragmatischer Aspekte zum Verstehen natürlicher Sprache. Dabei muß zusätzliches Wissen über die Struktur der domänenspezifischen Sprache miteingebracht werden. Bei den derzeit untersuchten Verfahren lassen sich die zwei kontrovers diskutierten Ansätze der reinen *Statistik* und der ausschließlich *regelbasierten Vorgehensweise* unterscheiden. Üblich sind mehrstufige Systeme, bei denen die signalnahen Verarbeitungsschritte zur Ermittlung der Wortketten mit stochastischen Methoden und die anschließende linguistische Analyse regelbasiert erfolgen. Die dabei auftretenden Defizite beim Abgleich der verwendeten Wissensbasen und der hohe Ressourcenaufwand erfordern jedoch eine noch engere Verzahnung der beiden Verfahren. Im hier vorgestellten Ansatz erfolgt die Decodierung des Bedeutungsinhaltes (dargestellt durch die semantische Gliederung) der Äußerung in einer einzigen Stufe mit Hilfe von zuvor trainierten stochastischen Wissensbasen, wie dies in Abbildung 4 dargestellt ist.

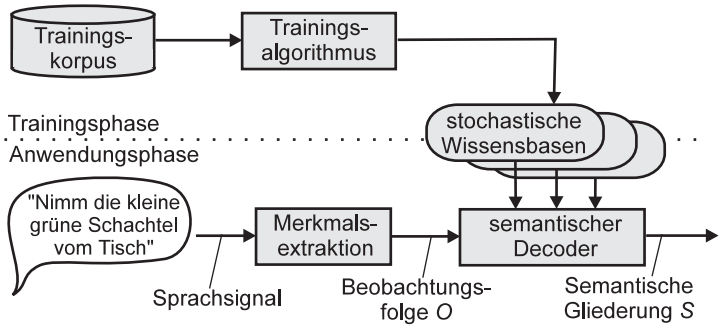

Abb. 4. Ermittlung des Bedeutungsinhalts mit stochastischen Methoden

Die Merkmalsextraktion erzeugt alle 10 ms einen 64-dimensionalen Merkmalsvektor, welcher die spektralen Eigenschaften des Sprachsignals innerhalb eines kurzen Zeitabschnittes beschreibt. Die zeitliche Abfolge solcher Merkmalsvektoren wird als Merkmalsvektorenfolge oder Beobachtungsfolge O bezeichnet. Sie dient als unmittelbare Eingabe-Symbolfolge für den semantischen Decoder.

3.1 Stochastische Maximum-a-posteriori-Klassifikation

Die semantische Decodierung bedient sich stochastischer Methoden, bei denen die menschliche Spracherzeugung als verrauschter Kanal zur Abbildung des Bedeutungsinhaltes S auf die Beobachtungsfolge O der Äußerung aufgefaßt wird.

Im gewählten Maximum-a-posteriori(MAP)-Ansatz wird nun diejenige semantische Gliederung S_{E} gesucht, welche die höchste Rückschluß(a-posteriori)-Wahrscheinlichkeit $P(S|O)$ zur gegebenen Beobachtungsfolge O aufweist. Diese Wahrscheinlichkeit läßt sich mit dem Satz von Bayes umformen:

$$S_{\mathrm{E}} = \underset{S}{\operatorname{argmax}} P(S|O) = \underset{S}{\operatorname{argmax}} \frac{P(O|S) \cdot P(S)}{P(O)} = \underset{S}{\operatorname{argmax}} [P(O|S) \cdot P(S)] \quad (1)$$

$P(O)$ wird bei der Maximierung nicht berücksichtigt, da sie bei gegebenem O konstant ist. Die direkte Bestimmung der Abbildungswahrscheinlichkeit $P(O|S)$ ist aufgrund der Vielfalt möglicher Kombinationen aus O und S nicht möglich. Das Modellierungsproblem für $P(O|S)$ wird daher in ebenenspezifische Teilprobleme überführt, indem die zusätzlichen Repräsentationsebenen Wortkette W und Lautfolge Ph eingeführt werden:

$$S_{\mathrm{E}} = \underset{S}{\operatorname{argmax}} \sum_{\text{alle } W} \sum_{\text{alle } Ph} [P(O|Ph) \cdot P(Ph|W) \cdot P(W|S) \cdot P(S)]. \quad (2)$$

Wird angenommen, daß nur die wahrscheinlichste Wortkette W und wahrscheinlichste Lautfolge Ph relevant sind, lassen sich die Summen zu Maximierungen über bedingte Wahrscheinlichkeiten vereinfachen:

$$S_{\mathrm{E}} = \underset{S}{\operatorname{argmax}} \max_{W} \max_{Ph} [P(O|Ph) \cdot P(Ph|W) \cdot P(W|S) \cdot P(S)] \quad (3)$$

Aus dieser wahrscheinlichsten Kombination einer semantischen Gliederung S, einer Wortkette W, einer Phonemkette Ph und der gegebenen Beobachtungsfolge O wird die erkannte semantische Gliederung S_{E} extrahiert. Dabei liefern vier stochastische Wissensbasen (semantisches, syntaktisches, phonetisches, akustisches Modell) die zur Maximierung benötigten bedingten Wahrscheinlichkeiten. Neu hierbei ist die Trennung von semantischem und syntaktischem Wissen [6], die eine MAP-Decodierung der semantischen Gliederung überhaupt erst ermöglicht.

Die stochastischen Wissensbasen modellieren die zur Decodierung benötigten Wahrscheinlichkeiten nicht als ganzes, sondern als Produkt mehrerer, als stochastisch unabhängig angenommener Wahrscheinlichkeiten. Das hat zum einen den Vorteil, daß die Anzahl der Parameter in den Modellen gering gehalten werden kann, und somit die Fähigkeit zum Generalisieren bereits bei einer kleinen Trainingsstichprobe gegeben ist. Zum anderen ermöglichen die Modelle die Realisierung eines sehr effektiven Algorithmus zur semantischen Decodierung, da sie in der Lage sind, die Wahrscheinlichkeiten zeitlich inkrementell in kleinen "Portionen" zu liefern [8].

3.2 Semantische Gliederung *S*

Als Schnittstelle zum nachfolgenden Befehlsgenerator dient die semantische Gliederung. Sie ist durch die folgenden Punkte charakterisiert [9]:

- Eine semantische Gliederung S ist ein Baum, bestehend aus N semantischen Untereinheiten (*Semunen*) $s_n : S = \{s_1, s_2, ..., s_n, ..., s_N\}$.
- Jedes Semun s_n besitzt einen Typ $t[s_n]$ und einen Wert $v[s_n]$.

– Jedes Semun s_n verweist auf eine bestimmte Anzahl $X \geq 1$ von sogenannten Nachfolger-Semunen $q_1[s_n], ..., q_X[s_n] \in \{s_2, ..., s_N, \text{leer}\} \setminus \{s_n\}$.
– Das leere Semun 'leer' steht für ein Blatt des Baumes. Es besitzt den Typ $t[\text{leer}] =$ 'leer', keinen Wert und keinen Nachfolger.

Abbildung 5 zeigt beispielhaft eine Wortkette W und die zugehörige semantische Gliederung S als unterschiedliche Repräsentationsformen einer Äußerung.

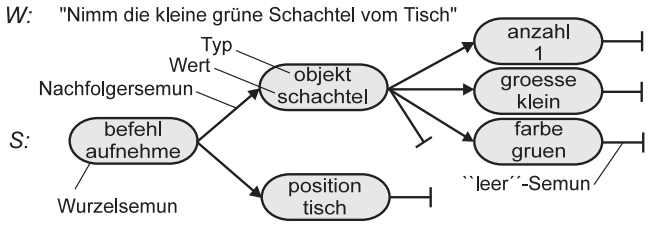

Abb. 5. Wortkette W und semantische Gliederung S

3.3 Bereitstellung der stochastischen Wissensbasen

Die Parameter des **semantischen und syntaktischen Modells** werden mit geeignetem Trainingsmaterial abgeschätzt. Zu jeder Äußerung werden Wortkette und semantische Gliederung mit Hilfe eines geeigneten Typ- und Werteinventars, welches den Bereich der zu erwartenden Bedeutungsinhalte abdeckt, gebildet und zum Training herangezogen. Für Äußerungen außerhalb der Domäne (z.B. "*heute ist schönes Wetter*") werden keine Typen und Werte definiert. Der derzeitige Trainigskorpus besteht aus 285 Sätzen, die von mehreren Sprechern gesammelt wurden. Bei der Auswahl der Sätze müssen folgende Punkte beachtet werden:

– Ansprechen jeder möglichen Funktion des Roboters
– Benennung aller möglichen und zukünftigen Serviceaufgaben
– Unterbringung aller Objekte, Orte und Positionen
– Abdecken der gewünschten quantitativen Einheiten (Abstände, Winkel, usw.)
– Verwendung verschiedener Worte und syntaktischer Konstrukte

Das **phonetische Modell** wurde als sogenanntes Transkriptionslexikon ausgeführt, das zu jedem Wort, welches beim Training der Grammatik verwendet wurde, jeweils die zugehörige Standardaussprache enthält. Erstellt wurde dieses Lexikon rein manuell unter Zuhilfenahme des Duden. Die **akustische Modellierung** wurde von einem bestehenden Spracherkennungssystem übernommen [10]. Das Training erfolgte mit domänenfremdem Trainingsmaterial, wobei durch Verteilung des Traingsmaterials auf 200 verschiedene Sprecher ein hohes Maß an Sprecherunabhängigkeit der Modellierung erzielt wurde [11].

4 Umsetzung in Roboterbefehle

Nach der Decodierung der semantischen Gliederung S generiert der Befehlsgenerator im nächsten Schritt den korrespondierenden Roboterbefehl, vergleiche Abbildung 2. Dieser planerverständliche Roboterbefehl, der sich an die in Abschnitt 2.3 definierte Befehlsstruktur hält, wird schließlich dem Serviceroboter übermittelt.

4.1 Anforderungen an die Umsetzung

Die Umsetzung der semantischen Gliederung in einen Roboterbefehl ist nicht immer ohne weiteres möglich. Es lassen sich verschiedene vom Nominalablauf abweichende Situationen erkennen: 1. Die vom Decoder generierte semantische Gliederung ist *nicht sinnvoll*. 2. Es fehlen für den Befehl wichtige Daten, d.h die semantische Gliederung ist zwar korrekt aber *unvollständig*. 3. Die in der semantischen Gliederung spezifizierten Daten, wie z.B. Objekte oder Räumlichkeiten, sind dem Gesamtsystem *nicht bekannt*. 4. Es handelt sich um einen *hochprioren* Überwachungsbefehl. Faßt man die verschiedenen Situationen zusammen, so lassen sich die folgenden Anforderungen an den Befehlsgenerator ableiten:

- Übersetzung der semantischen Gliederungen in Roboterbefehle
- Validierung der Befehle auf Plausibilität und Vollständigkeit
- Einbindung aktueller Umgebungsinformation und des Roboterzustandes
- Dialoggeführtes Einbeziehen des Bedieners zur Vervollständigung der Befehle unter Ausnutzung der zur Verfügung stehenden MRI-Ressourcen
- Einfache Anpassung an neue Eingabecodes bzw. eine neue Domäne

4.2 Befehlsgenerator

Der in Abbildung 6 gezeigte Befehlsgenerator wandelt die semantische Gliederung schrittweise in die Struktur des Befehlsstrings um. Durch auftretende Redundanz oder Mehrdeutigkeit kann die Ersetzung nicht ohne weiteres erfolgen. Oftmals muß der Nachfolger eines Semuns bekannt sein. Die in einem Umsetzmodell zusammengefaßte Ersetzungsstrategie bietet somit die Möglichkeit, Ersetzungen bedingt durchzuführen oder Datenelemente direkt zu übernehmen. Mit Hilfe des Umsetzmodells wird sowohl die Konsistenzprüfung während der Ersetzung als auch die Vollständigkeitsuntersuchung durchgeführt. Bei unvollständigen oder mehrdeutigen Bedienereingaben wird der Befehl nicht verworfen, sondern durch eine Rückfrage an den Bediener ergänzt. Anschließend erfolgt eine erneute Prüfung. Mit Hilfe der statischen und dynamischen Wissensbasen wird die Existenz der verwendeten Objekte und Räumlichkeiten überprüft. Die mit Namen vorhandenen Objekte und Räumlichkeiten werden dabei zusätzlich durch ihre kartesische Position ergänzt. Bei Nichtvorhandensein oder Mehrdeutigkeiten wird der Bediener nochmals zur genaueren Spezifikation eingeschaltet. Im letzten Schritt erfolgt für einen Makrobefehl die Expansion in von der Planungsebene unterstützte Aufträge. Mit der Möglichkeit, Rückfragen an den Bediener zu stellen, steht ein leistungsstarker Befehlsgenerator zur Verfügung, der einen nicht ganz verstandenen Befehl nicht sofort verwirft, sondern gezielt nachhakt.

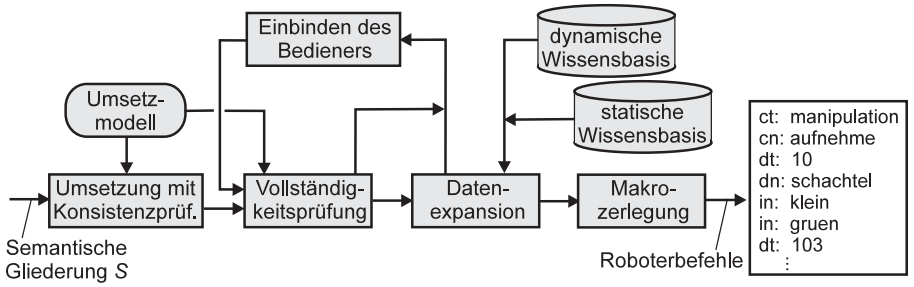

Abb. 6. Befehlsgenerator

4.3 Einsatz weiterer MRI-Ressourcen

Nicht immer ist es ausreichend bzw. zweckmäßig, den Servicerobter ausschließlich über die Spracheingabe zu kommandieren. Weitere, einfach zu bedienende Ressourcen sollten auf Seiten des MRI dem Bediener bereitgestellt werden. Wesentlich einfacher ist es z.B., mit dem Finger auf ein Objekt zu zeigen oder eine kartesische Position zu spezifizieren. Dies kann durch Anklicken des interessierenden Objektes in einer das Arbeitsumfeld darstellenden virtuellen Welt per Computermaus oder Touchscreen erfolgen. Aber auch die präzise Telemanipulation des Roboters durch Vorgabe kleinster Relativbewegungen ist mittels sprachlicher Steuerung umständlich und langwierig. Entsprechende Eingabeelemente für Kraft und Position werden zusätzlich bereitgestellt. Auf der Ausgabeseite kommen ebenfalls unterschiedliche Ressourcen zum Einsatz. Zur bildschirmbasierten Kommandierung, Überwachung und Unterstützung dient das Bild des virtuellen Arbeitsumfelds und das reale Videobild von einer onboard-Kamera. Meldungen erscheinen in einem Textfenster oder werden über eine Sprachausgabe an den Bediener weitergeleitet.

Es ist wichtig, eine leistungsstarke und transparente Verwaltungsstruktur für die unterschiedlichen MRI-Ressourcen bereitzustellen. Sie stellt sicher, daß anforderungsabhängig eine gezielte Einbindung unterschiedlicher Ein- und Ausgabemedien in das MRI möglich wird. Die Ressourcen müssen sich dabei über die Verwaltungsstruktur zu einem Gesamtsysten zusammenfügen, das es dem Bediener erlaubt, einfach und schnell mit dem Roboter zu kommunizieren.

5 Experimentelle Ergebnisse

5.1 Experimentierplattform ROMAN und MRI

Mit der Experimentierplattform ROMAN steht ein semiautonomer Serviceroboter für Innenraumumgebungen zu Verfügung, siehe Abbildung 1 [7]. ROMAN besteht mit seiner kompakten Bauweise aus dem Mehrgelenkmanipulator MANUS, erweitert durch eine Hoch/Tief-Linearachse, und einer sehr wendigen Plattform, welche den Arbeitsraum des Manipulators gezielt erweitert oder reine Transportaufgaben ermöglicht. Die verwendete Sensorik besteht aus einer der absoluten Positionsbestimmung dienenden Lasernavigationseinheit mit kreiselbasierter Koppelnavigation, dem Ultraschallring zur Hindernisvermeidung und einer Videosensorik zur Objekterkennung und -verfolgung. In der Planungsebene des Roboters stehen Grundfunktionalitäten zur Verfügung, welche ein flexibles und situationsabhängiges Verhalten ermöglichen [12].

Mit dem MRI ist ROMAN über ein Funkethernet und eine Hochfrequenz-Videostrecke verbunden. Das MRI basiert auf einer Grafikworkstation mit eingebautem Videoboard zur online-Darstellung der Kamerabilder. Das Virtual-Reality-Tool AnySIM stellt die aktuelle Roboterposition und Konfiguration im virtuellen Arbeitsraum dar. Meldungen erfolgen über eine Sprachausgabe an den Bediener oder an Personen im Arbeitsraum. Zur Eingabe von Kommandos dient das in diesem Beitrag vorgestellte System zum Verstehen natürlicher, gesprochener Sprache. Das dazu aufgestellte Trainingsmaterial besteht derzeit aus 285 Wortketten, das daraus ableitbare Vokabular enthält 409 Wörter. Es werden 42 Typen und 214 Werte unterschieden. Weitere Eingabemedien sind eine Spacemouse zur Unterstützung von Operationen und die Computermaus zur Auswahl von Objekten und Positionen im virtuellen Arbeitsraum [13].

5.2 Experimente

Im ersten Experiment gemäß Abbildung 7 fordert der Bediener den Serviceroboter auf, ihm die Kaffeetasse vom seinem Schreibtisch zu holen. Die generierte semantische Gliederung spezifiziert die Serviceaufgabe vollständig. Als Ziel ist der den Auftrag gebende Bediener eingesetzt. Nach der Umsetzung werden die Datenblöcke 2 und 3 noch um die korrespondierende kartesische Position ergänzt. Dies geschieht durch einen Blick in die statische bzw. dynamische Wissensbasis auf Seiten des MRI. Bevor der Makrobefehl an den Roboter geschickt wird, erfolgt eine Zerlegung in die beiden Teilaufträge *aufnehmen* und *abstellen*. Daraufhin führt der Roboter den Auftrag aus.

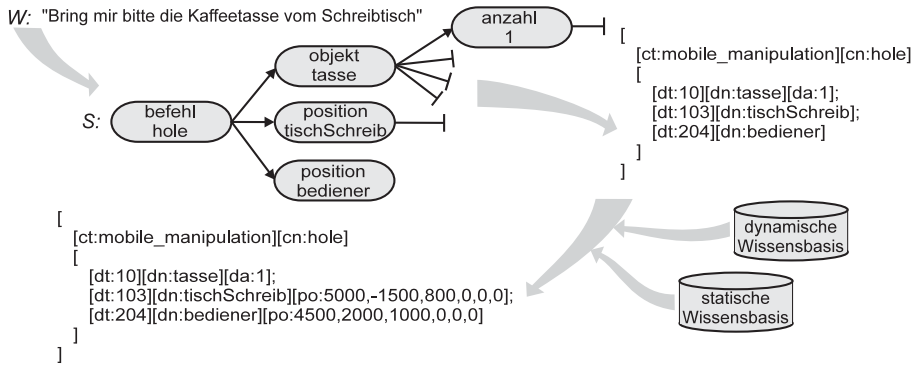

Abb. 7. Experiment ohne Rückfrage an den Bediener

Ausgangspunkt des in Abbildung 8 dargestellten zweiten Experiments ist der unvollständige Befehl, die Kaffeetasse wegzuräumen. Die Vollständigkeitsprüfung ergibt zwei fehlende Datenblöcke, zum einen die Objektlage und zum anderen die Zielposition. Der Bediener wird daraufhin eingebunden und spezifiziert die fehlenden Daten durch Anklicken in der virtuellen Welt. Der ergänzte Makrobefehl wird dann zerlegt und zur Ausführung an den Planer des Roboters weitergeschickt.

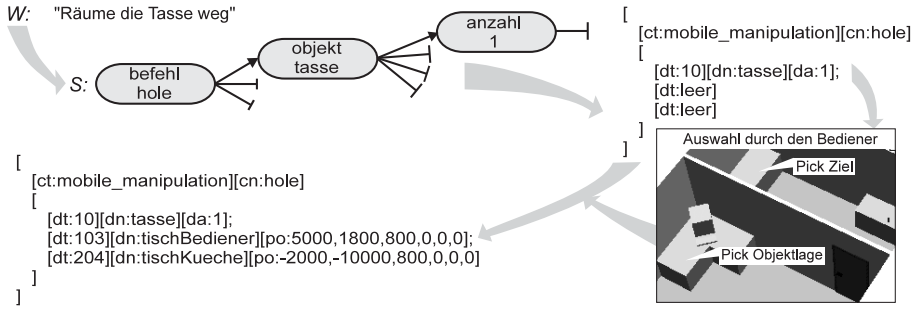

Abb. 8. Experiment mit Rückfrage an den Bediener

5.3 Bewertung

Für die semantische Decodierung von Äußerungen, die vollständig im Trainingskorpus der Grammantik enthalten sind, wurde in [8] eine Trefferrate von 75%

erzielt, allerdings über der Domäne eines Grafikeditors mit ähnlicher Anzahl von Typen und Werten, aber doppelter Vokabulargröße (854 Wörter). Es ist zu erwarten, daß aufgrund des nur halb so großen Vokabulars der Roboterdomäne deutlich weniger Verwechslungen von Wörtern auftreten, was sich positiv auf das Ergebnis auswirkt.

Äußerungen von Benutzern, die nicht mit dem System vertraut sind, werden in der Regel nicht im Trainingskorpus enthalten sein. Probleme bereitet eine solche Äußerung prinzipiell dann, wenn sie neue, unbekannte Wörter aufweist. Dieser sogenannte OOV(Out-of-Vocabulary)-Fall ist unerwünscht, resultiert er doch meist in einem Fehler bei der semantischen Decodierung. Nach [14] läßt sich die OOV-Rate für das verwendete Trainingsmaterial zu 53% abschätzen. Um die OOV-Rate auf unter 10% zu senken, müßte der Trainingskorpus um 8.800 weitere Wortketten erweitert werden. Das resultierende Vokabular umfaßt dann etwa 2.915 Worte.

Der Befehlsgenerator setzt richtig verstandene, womöglich unvollständige semantische Gliederungen immer korrekt in einen Roboterbefehl um. Dazu können bei der Vervollständigung sowohl alle MRI-Ressourcen als auch der Bediener gefordert sein. Falsch decodierte semantische Gliederungen können zum großen Teil durch Konsistenzprüfung erkannt und somit verworfen werden. Um aber auch die Befehle zu erkennen, die zwar falsch sind, aber sich korrekt in einen Roboterbefehl übersetzen lassen, muß zur Gewährleistung 100%-iger Sicherheit immer eine Rückfrage an den Bediener erfolgen.

6 Zusammenfassung und Ausblick

Das hier vorgestellete Konzept bildet einen Beitrag zur einfachen, dem Menschen vertrauten Kommandierung eines semiautonomen Serviceroboters. Grundlage dafür ist die Verwendung sprecherunabhängiger, natürlicher, gesprochener Sprache. Basierend auf der Definition der domäneneigenen Sprache konnten die Sprachmodelle generiert werden. Der domänenunabhängige semantische Decoder stellt mit der semantischen Gliederung eine definierte Schnittstelle zu der sich anschließenden Befehlsgenerierung dar. Bei der Umsetzung in Roboterbefehle werden falsche semantische Gliederungen abgefangen und unvollständige Befehle durch das Einbinden des Bedieners ergänzt. Der resultierende Roboterbefehl genügt einer Befehlsstruktur, in der alle zur Verfügung stehenden Informationen dem Planer des Roboters geordnet bereitstellt werden. Der durchgängige Ansatz von einem akustischen Signal bis hin zum korrespondierenden Roboterbefehl ist durch hohe Flexibilität, Sicherheit und Robustheit gekennzeichnet. Das System zum Verstehen natürlicher, gesprochener Sprache ist in das bestehende MRI integriert und dient der Kommandierung des Serviceroboters ROMAN.

Zukünftige Arbeitsschwerpunkte werden sich mit der Erweiterung des Wortschatzes der Domäne und der Verwendung englischer Sprache beschäftigen. Ferner ist daran gedacht, weitere Umgebungsinformationen mit in die Umsetzung einzubinden.

Danksagung

Der vorliegende Beitrag ist das Ergebnis einer Zusammenarbeit zwischen dem Lehrstuhl für Steuerungs- und Regelungstechnik und dem Lehrstuhl für Mensch-Maschine-Kommunikation, beide TU München. Das Projekt ROMAN wird im

Rahmen des Sonderforschungsbereichs *Informationsverarbeitung in autonomen, mobilen Handhabungssystemen* (SFB 331) von der Deutschen Forschungsgemeinschaft (DFG) gefördert.

Literatur

1. K. Kawamura, M. Iskarous, "Trends in service robots for the disabled and the elderly", in *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, München, Deutschland, 1994, S. 1647–1654.
2. V. Dahl et al., "Driving robot through natural language", in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Vancouver, Canada, 1995, S. 1904–1908.
3. B. Caprile, G. Lazarri, "Commanding a robot by voice: Speech and autonomous navigation for the mobile robot of MAIA", in *Robotics in Alpe-Adria Region*, Springer Verlag, Wien, New York, 1994, S. 153–157.
4. E. Stopp, Th. Laengle, "Natürlichsprachliche Instruktionen an einen autonomen Serviceroboter", in *Tagungsband zum 11. Fachgespräch Autonome Mobile Systeme*, Karlsruhe, Deutschland, R. Dillmann, U. Rembold, T. Lüth (Hrsg.), Springer Verlag, 1995, S. 299–308.
5. S. Förster, K. Peters, "Sprachliche Steuerung behaviourorientierter Systeme", in *Tagungsband zum 11. Fachgespräch Autonome Mobile Systeme*, Karlsruhe, Deutschland, R. Dillmann, U. Rembold, T. Lüth (Hrsg.), Springer Verlag, 1995, S. 309–318.
6. H. Stahl, J. Müller, "A stochastic grammar for isolated representation of syntactic and semantic knowledge", in *Proceedings of the European Conference on Speech Communication and Technology*, Madrid, Spanien, 1995, S. 551–554.
7. W. Daxwanger et al., "ROMAN: Ein mobiler Serviceroboter als persönlicher Assistent in belebten Innenräumen", in *Tagungsband zum 12. Fachgespäch Autonome Mobile Systeme*, München, Deutschland, G. Schmidt, F. Freyberger (Hrsg.), Springer Verlag, 14.–15. Oktober 1996.
8. H. Stahl, J. Müller, M. Lang, "An efficient top-down parsing algorithm for understanding speech by using stochastic syntactic and semantic models", in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Atlanta, USA, 1996, S. 397–400.
9. J. Müller, H. Stahl, "Die semantische Gliederung als adäquate semantische Repräsentation für einen sprachverstehenden Grafikeditor", in *Angewandte Computerlinguistik, Reihe "Sprache und Computer"*, Band 15, Georg Olms Verlag, Hildesheim, Deutschland, 1995, S. 211–225.
10. X.D. Huang et al., "Hidden markov models for speech recognition", in *Edinburgh University Press*, Edinburgh, England, 1990.
11. B. Pompino et al., "PhonDat Datenformate", in *Forschungsbericht Nr. 30, Institut für Phonetik und sprachliche Kommunikation, Ludwig-Maximilians-Universität*, München, Deutschland, 1992.
12. C. Fischer, M. Buss, G. Schmidt, "Soft control of an effector path for a mobile manipulator", in *Proceedings of the International Symposium on Robotics and Manufacturing (ISRAM)*, Montpellier, Frankreich, 1996.
13. C. Fischer, M. Buss, G. Schmidt, "Hierarchical supervisory control of service robot using human-robot-interface", in *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, Osaka, Japan, 4.–8. November 1996.
14. J. Müller, H. Stahl, M. Lang, "Predicting the out-of-vocabulary rate and the required vocabulary size for speech processing applications", in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, 3.–6. Oktober 1996.