Aus *Künstliche Intelligenz: Themenheft Mustererkennung*, Band 3/95, pp. 35–41

Repräsentation und Nutzung von Wissen in einem Spracherkennungs- und Dialogsystem

H. Niemann, G. Sagerer, F. Kummert, M. Mast

1 Einführung

Das Ziel der *automatischen Spracherkennung* ist die zuverlässige, sprecherunabhängige und schnelle Erkennung der in einer Äußerung enthaltenen Wörter, die aus einem genügend großen Vokabular sind. Arbeiten in dieser Richtung sind z.B. in [Bah83, Kun91, Low76, Rab88, ST94] zu finden.

Eine linguistische Analyse wird in der *Verarbeitung natürlicher Sprache* betrieben, die sich auf die Analyse gedruckter Texte konzentriert. Dabei stehen Probleme der Ermittlung der syntaktischen Struktur und der semantischen und pragmatischen Eigenschaften im Vordergrund. Die bei der Verarbeitung gesprochener Sprache auftretenden Probleme der falsch erkannten oder ausgelassenen Wörter oder der Äußerungen, die keinen vollständigen Satz bilden, können dann unbeachtet bleiben. Beispiele für Arbeiten in dieser Richtung sind in [All89, Hir89, Wah88].

Systeme, die zu einer gesprochenen Äußerung eine Antwort oder eine Übersetzung generieren, müssen das Gesagte zunächst *verstehen*, d.h. in ein systeminternes Schema zur Repräsentation aufgabenspezifischen Wissens einordnen. Wenn die Worterkennung fehlerfrei ist, kann man im Prinzip daran denken, die Ausgabe eines Spracherkenners als Eingabe eines natürlichsprachlichen Systems zu verwenden. Wegen Unsicherheiten bei der Worterkennung und Unterschieden zwischen geschriebenen Texten und fließender Rede ist dieses jedoch nicht möglich. Erforderlich ist eine Systemarchitektur, in der eine Interaktion zwischen Worterkennung und linguistischer Analyse möglich ist und in der letztere trotz unvollständiger Äußerungen und fehlerhafter Worterkennung arbeiten kann. Arbeiten zu sprachverstehenden Systemen findet man z.B. in [Nie88, You89, Bat93].

In diesem Beitrag orientieren wir uns an dem Spracherkennungs- und Dialogsystem EVAR, dessen Struktur in Bild 1 gezeigt ist. Das System folgt dem geschichteten linguistischen Modell, ohne jedoch an eine strikte stufenweise (daten- oder modellgetriebene) Verarbeitungsstrategie gebunden zu sein. Die wesentlichen Verarbeitungsschichten gehen aus dem Bild hervor. In einer ersten Verarbeitungsphase werden aus dem Sprachsignal Wörter ermittelt. Diese bilden die Eingabe für die linguistische Verarbeitung in der zweiten Phase, die aus den Teilen Syntax, Semantik, Pragmatik, Dialog und Antwortgenerierung besteht. Die Schnittstelle zwischen beiden Verarbeitungsphasen ist also die Ebene der Wörter. Die Verarbeitungsstrategie wird von einem Kontrollmodul in Abhängigkeit von gespeichertem Wissen und dem Sprachsignal bestimmt.

Dieser Beitrag konzentriert sich auf Arbeiten, die im Rahmen von zwei Einzelvorhaben des DFG-Schwerpunktprogramms durchgeführt wurden. Es sind die Vorhaben:

1. Entwicklung eines Kontrollalgorithmus mit zugehörigen Bewertungsfunktionen für ein wissensbasiertes System mit mehreren unabhängigen Modulen zum Führen von Dialogen in kontinuierlich gesprochener Sprache [Kum91].

2. Entwicklung und Realisierung eines Dialogmoduls für ein System zum Verstehen kontinuierlich gesprochener deutscher Sprache [Mas93].

Das System EVAR in der gegenwärtigen Realisierung beruht auf weiteren Arbeiten, die im Rahmen von Verbundvorhaben durch das BMFT gefördert wurden und sich auf Worterkennung, Prosodie, Semantik und Pragmatik konzentrierten. Einzelheiten dieser Arbeiten finden sich in [Bri84, Bri87, Ehr86, Ehr90, Kun91, Nöt91, ST87].

Die beschriebene Systemarchitektur hat die folgenden wesentlichen Eigenschaften:

- Allgemeines linguistisches und aufgabenspezifisches Wissen werden in wohlstrukturierter Weise repräsentiert, so daß
  - ein einheitlicher Formalismus zur Wissensrepräsentation in allen Schichten des Systems möglich ist,
  - beliebiges prozedurales Wissen integriert werden kann,
  - die Analyse von mehr als einer Äußerung möglich ist, d.h. es kann ein Dialog zwischen System und Benutzer erfolgen.
- Gleichzeitig ist eine flexible Kontrollstrategie möglich, die
  - zwischen Phasen datengetriebener und modellgetriebener Verarbeitung alterniert,
  - auf einem theoretisch motivierten Ansatz zur Erreichung eines globalen Optimums des Systemverhaltens beruht,
  - die Integration lokaler Heuristiken zur Beeinflussung der Verarbeitungsstrategie erlaubt,
  - eine Interaktion zwischen Worterkennung und linguistischer Verarbeitung ermöglicht.

Die Aufgabe des erwähnten Spracherkennungs- und Dialogsystems ist die Beantwortung von Anfragen nach Intercity-Zugverbindungen. Das System soll insbesondere in der Lage sein, auf unvollständige oder unverstandene Fragen des Benutzers mit gezielten Rückfragen zu antworten, also einen sinnvollen und zielgerichteten Dialog zu führen. Dafür sind die Teilaufgaben Erkennen der gesprochenen Wörter, Verstehen der Äußerung, Generieren einer Antwort oder Generieren einer Rückfrage zu bewältigen, wovon das Acronym EVAR abgeleitet ist. Das Vokabular liegt bei ca. 1000 Wörtern. Die Syntax soll hinreichend allgemein sein, so daß der Sprecher keine besonderen Einschränkungen beachten muß. Die Worterkennung soll sprecherunabhängig erfolgen, jedoch werden Dialekte ausgeschlossen. Systemüberblicke sind in [Nie85, Nie88, Mas94] gegeben.

In diesem Beitrag wird zunächst die linguistische Verarbeitung erörtert. Dabei wird exemplarisch die oberste

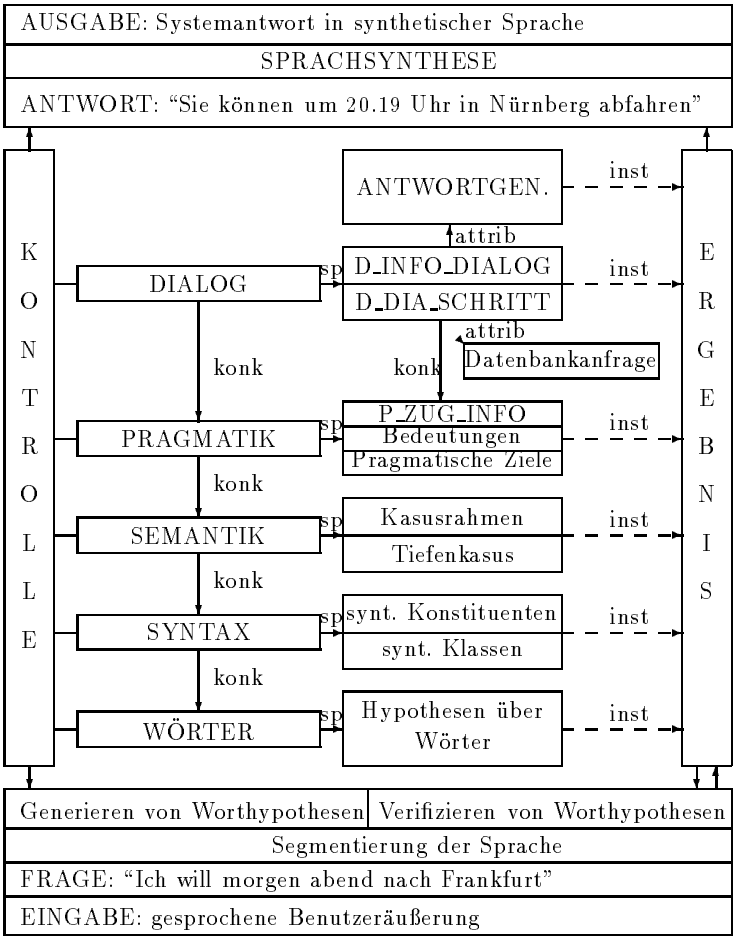

Abbildung 1: Eine Übersicht über das Spracherkennungs- und Dialogsystem EVAR

Ebene (Dialog) behandelt, während Syntax, Semantik und Pragmatik nur kurz skizziert werden. Ein entscheidender Aspekt der Analyse ist die Fähigkeit, alternative Interpretationen des Sprachsignals auf allen Verarbeitungsebenen zu bewerten; darauf wird in Abschnitt 3 eingegangen. Erzielte Ergebnisse werden in Abschnitt 4 vorgestellt.

2 Linguistisches Wissen

Sprachverstehen wird als eine Folge von Operationen aufgefaßt, die das Sprachsignal über verschiedene Abstraktionsebenen in die gewünschte Ausgabegröße — hier eine Antwort oder im allgemeinen eine sinnvolle Systemreaktion auf eine Benutzeranfrage — transformieren. Die Operationen sind abhängig von dem eingegebenen Sprachsignal, erzielten Zwischenergebnissen und gespeichertem Wissen. Als Repräsentationsformalismus wird die semantische Netzwerksprache ERNEST [Sag90, Kum91] verwendet. Da diese an anderer Stelle in diesem Heft beschrieben wird, wird hier nur kurz darauf eingegangen. Desweiteren wird erläutert, welches Wissen in den fünf Ebenen 'WÖRTER — DIALOG' in Bild 1 zur Analyse des Sprachsignals zur Verfügung steht.

2.1 Repräsentationsformalismus

Zur Repräsentation des linguistischen und anwendungsabhängigen Wissens und zur Modellierung des Dialogs dient eine homogene Wissensbasis, die als semantisches Netz in ERNEST realisiert ist. In dieser Netzsprache sind drei Knotentypen (Konzept, Instanz, modifiziertes Konzept) und drei Kantentypen (Spezialisierung, Bestandteil, Konkretisierung) fest definiert. Ein **Konzept** dient der Modellierung eines Begriffs, einer Objektklasse oder einer Klasse von Ereignissen. In der hier beschriebenen Anwendung sind dies beispielsweise Begriffe wie Nomen, Präpositionalgruppe, Tiefenkasus, Verbrahmen usw. Um das Sprachsignal mit den Begriffen des Problemkreises interpretieren zu können, werden Signalausschnitte mit Konzepten und damit mit deren Bedeutung verbunden. Eine solche Verbindung wird über eine **Instanz** etabliert, wobei eine Instanz immer genau *einem* Konzept zugeordnet ist. Diese Zuordnung führt in vielen Fällen zu Einschränkungen für die Interpretation des restlichen Signals. Um diese Restriktionen darstellen zu können, wurde das **modifizierte Konzept** eingeführt. Es repräsentiert somit Wissen, das an eine konkrete Analysesituation adaptiert wurde. Durch den Kantentyp **Spezialisierung** werden Konzepte als Ober- und Unterbegriffe miteinander verbunden, z.B. Wortart $\stackrel{spez}{\rightarrow}$ Nomen. Die Beziehung, daß ein Konzept Teil eines anderen ist, wird durch den Kantentyp **Bestandteil** ausgedrückt, z.B. Nominalgruppe $\stackrel{bst}{\rightarrow}$ Nomen. Diese reine Bestandteilbeziehung gilt jedoch nicht überall. So ist zwar ein bestimmter Tiefenkasus Bestandteil eines Verbrahmens, aber die Konstituente "der Zug" kann nur als Tiefenkasus *Instrument* interpretiert werden, falls sie im Kontext eines bestimmten Verbrahmens (z.B. "abfahren") auftritt. Im Satz "Hat der

Zug einen Speisewagen" besitzt die Konstituente "der Zug" den Tiefenkasus *Object*. Deshalb ist *Instrument* ein kontextabhängiges Bestandteil des Verbrahmens "abfahren" (S_VR_ABFAHREN $\stackrel{kbst}{\rightarrow}$ S_INSTRUMENT). Konzepte, die Begriffe aus unterschiedlichen Abstraktionsebenen darstellen, werden über den Kantentyp **Konkretisierung** in Beziehung gesetzt, z.B. Tiefenkasus $\stackrel{kon}{\rightarrow}$ Nominalgruppe.

2.2 Wörter

Das System in Bild 1 arbeitet in den zwei Phasen der datengetriebenen Worterkennung und der modellbasierten linguistischen Analyse. Der Grund dafür ist, daß eine wissensbasierte Verarbeitung nicht Abtastwerte des Sprachsignals als Eingabe verwendet sondern grössere Einheiten; diese sind hier Wörter, jedoch sind z.B. auch syntaktische Konstituenten als abstraktere oder Silben als konkretere, d.h. näher am Sprachsignal liegende, Einheiten denkbar.

Das Sprachsignal wird mit 16 kHz abgetastet, in nichtüberlappende Sprachrahmen von 10 ms Dauer zerlegt und durch 24 cepstrale Merkmale, die gehörbezogen berechnet werden und auch zeitliche Änderungen berücksichtigen, parametrisch repräsentiert. Ein Normalverteilungsklassifikator ordnet jedem Sprachrahmen bis zu fünf Phonkomponenten und deren a posteriori Wahrscheinlichkeit zu. Die klassifizierten Sprachrahmen werden mit 'Hidden Markov Modellen' (HMM) zu lautlichen Segmenten zusammengefaßt. Aus diesen werden wiederum mit HMM, die automatisch aus der Standardaussprache eines Lexikons generiert werden, *Worthypothesen* berechnet. Die Ergebnisse dieser Verarbeitungsphase werden durch die *Schnittstellenebene* WÖRTER an die linguistische Verarbeitung übergeben.

Da die Worterkennung fehlerhaft ist, werden erheblich mehr Worthypothesen generiert und weitergegeben als Wörter im Sprachsignal vorhanden sind, um mit hoher Wahrscheinlichkeit viele der tatsächlich gesprochenen Wörter in der Menge der Worthypothesen zu finden. Jede Worthypothese hat eine Bewertung, die von der Ausgabewahrscheinlichkeit des HMM abgeleitet wird. Wie in Abschnitt 3 ausgeführt wird, bilden diese eine wichtige Grundlage für die Bewertung von Alternativen in der linguistischen Verarbeitung.

Die linguistische Analyse setzt Wörter aus der Menge der Worthypothesen zu längeren, im Sinne des gespeicherten Wissens korrekten oder *kompatiblen* Wortketten zusammen. Die in Bild 1 gezeigte Wortverifikation erlaubt es, solche neu gebildeten Wortketten nachträglich wieder mit dem Sprachsignal zu vergleichen und neu zu bewerten. Die Interaktion zwischen Worterkennung und linguistischer Analyse beruht zur Zeit auf dieser Neubewertung.

Die dem System bekannten Wörter sind in einem *Lexikon* repräsentiert, das außer der Standardaussprache eines Wortes auch Information zu dessen syntaktischen, semantischen, pragmatischen und dialogischen Eigenschaften enthält. Das Lexikon enthält Vollformen, d.h. neben der Grundform eines Wortes auch alle Flexionen dazu, da diese für die Worterkennung gebraucht werden.

In der gegenwärtigen Version des Systems EVAR wird zur Worterkennung ein Lexikon mit 1081 Wörtern benutzt; je Äußerung werden z.B. 50 – 100 Worthypothesen oder die beste Kette an die linguistische Verarbeitung weitergegeben. Daraus ergibt sich die wichtige Forderung, daß diese mit einer *Menge von Worthypothesen* arbeiten können muß. Eine Worthypothese w_{Hyp} ist ein Tupel

$$w_{Hyp} = (Güte, Wortnummer, Anfangszeit, Endzeit)\,.$$

Die Einzelheiten der Lexikonstruktur und der obigen Verarbeitungsschritte sind in [Ehr86, Kun91, ST94] enthalten.

2.3 Syntax

Im System EVAR wird eine strikte Zerlegung des linguistischen Wissens in die Ebenen 'Syntax, Semantik und Pragmatik' vorgenommen. Dazu kommen weiterhin die Ebene 'Dialog' zur Analyse einer Äußerung *im Kontext* eines Gesprächs und die Antwortgenerierung, in der nach dem Verstehen einer Äußerung und der Bereitstellung der Fakten für eine Antwort diese als deutscher Satz formuliert wird. Der Grund für diese Zerlegung ist die klare konzeptuelle Struktur des Systems sowie die weitgehende Unabhängigkeit der Schichten, was der Wartbarkeit und Wiederverwendbarkeit zugute kommt. Wird beispielsweise ein anderer Problemkreis betrachtet, so ist im wesentlichen die Pragmatik zu ändern, aber nicht die Syntax und Semantik.

Unter *Syntax* werden hier die Eigenschaften und Relationen von Wörtern und Wortklassen verstanden. Semantische oder andere Information wird auf dieser Ebene ausdrücklich nicht repräsentiert und nicht genutzt. Andererseits sollte semantische Information zur Reduzierung der Zahl der Alternativen bereits zu einem frühen Stadium der Analyse genutzt werden können. Daraus resultiert die Forderung nach einer *Analysestrategie*, die nicht strikt sequentiell von Ebene zu Ebene fortschreitet (gleichgültig ob dieses modellgetrieben ('top-down') oder datengetrieben ('bottom-up') geschieht), sondern die ein Alternieren zwischen verschiedenen Ebenen zuläßt. Dieses ist durch die Einführung eines separaten Kontrollmoduls möglich, der eine Folge von Verarbeitungsschritten berechnet.

Die Syntaxanalyse in einem Spracherkennungs- und Dialogsystem sollte neben der eigentlichen Aufgabe, nämlich der Bestimmung der syntaktischen Stuktur einer Äußerung, auch zur Reduktion der Unsicherheiten bei der Worterkennung beitragen und bei unvollständigen Äußerungen wenigstens noch ein Teilergebnis liefern. Solche Unvollständigkeiten können durch Unachtsamkeit des Sprechers (Auslassung von Wörtern, Verschleifung von Endungen), durch übliche Reduktionen von Antworten im Dialogkontext (insbesondere elliptische und anaphorische Wendungen) sowie durch Fehler in der Worterkennung verursacht sein.

Daher werden Äußerungen syntaktisch in Konstituenten zerlegt. Größere Einheiten werden auf dieser Ebene nicht betrachtet, da sich dies bei der Verwendung einer Menge von Worthypothesen als günstig erwiesen hat. Die Syntaxanalyse nimmt also nur die Berechnung der einfachen syntaktischen Konstituenten Nominalgruppe, Präpositionalgruppe, Verbalgruppe, prädikative Adjektivgruppe, Adverbialgruppe und Zeitangabe vor. Als 'einfach' werden Konstituenten wie 'mit dem nächsten Zug' oder 'nach Hamburg' bezeichnet, nicht dagegen Konstituenten wie 'der nächste Zug nach Hamburg', die erst in der semantischen Analyse gebildet werden. Damit ist es möglich, auch unvollständige Äußerungen, die aber syntaktisch korrekt gebildete Konstituenten enthalten, zu analysieren. Die Bildung von vollständigen Sätzen wird, wenn möglich, erst in der semantischen Analyse versucht.

Um die Dialogführung zu unterstützen, wurde auf der Syntaxebene die Behandlung einiger Redewendungen wie z.B. 'vielen Dank' vorgesehen.

Das syntaktische Wissen wird, wie oben erwähnt, in einem semantischen Netz repräsentiert. Eine Konsequenz dieser Darstellung ist, daß es keinen separaten Parser gibt, da das Auffinden syntaktischer Konstituenten der Instantiierung des zugehörigen Konzepts entspricht und die Instantiierung vom Kontrollalgorithmus gesteuert wird.

2.4 Semantik und Pragmatik

Die semantische Analyse wird hier nur kurz erwähnt. Sie beruht auf der Kasus- und Valenztheorie [Abr78, Fil68, Tes66] und verläuft in den folgenden drei Schritten:

1. Die von der Syntaxanalyse gelieferten syntaktischen Konstituenten werden auf *semantische Konsistenz* geprüft. Zur Effizienzsteigerung wird in diesem Schritt auch die pragmatische Konsistenz geprüft. Es gibt Tests für
   - die Konsistenz von Präposition und Adjektiv mit dem zugehörigen Nomen, wodurch Konstituenten wie 'der schnelle Baum' eliminiert werden;
   - ein Nomen im Singular ohne Artikel, wodurch Konstituenten wie 'was kostet Fahrkarte' eliminiert werden;
   - den Widerspruch semantischer Eigenschaften, wodurch Konstituenten wie 'ein nächster Zug' eliminiert werden.
2. Es werden *komplexe Konstituenten* konstruiert. Dazu gehören insbesondere Konstituenten wie 'der Zug nach Frankfurt' oder 'morgen früh gegen acht Uhr'. Es werden auch, wenn möglich, Satzhypothesen unter Einbezug eines Verbs bestimmt.
3. Wenn möglich wird eine Satzhypothese im semantischen Netzwerk instantiiert, wobei auch freie Angaben mit berücksichtigt werden.

Die pragmatische Analyse basiert auf einem Modell des Problemkreises, in dem hier die wesentlichen Typen von Zugauskünften und die dafür erforderlichen Elemente definiert werden. So erfordert z.B. eine Fahrplanauskunft vom Benutzer u.a. die Angabe einer ungefähren Abfahrts- oder Ankunftszeit, um die Zahl der in Frage kommenden Züge zu begrenzen. Wenn in einer Anfrage obligatorische Elemente für die Generierung einer Antwort fehlen, kann aufgrund des Modells vom System eine gezielte Rückfrage ermittelt werden. Einzelheiten dieser Verarbeitungsschritte sind in [Bri84, Ehr90] zu finden.

2.5 Dialog

Das System arbeitet in einem begrenzten Problemkreis. Der Benutzer möchte eine Information über Intercity-Zugverbindungen haben, d.h. es liegt ein Auskunftsdialog vor und nicht ein Verkaufsgespräch oder anderes. Der Dialog soll über Telefon erfolgen, so daß nichtverbale Kommunikation ausscheidet, dagegen wechselseitige Bestätigungen häufig sind. In spontaner Sprache treten 'Nicht-Wörter' (wie Husten, Räuspern), Satzbrüche und Häsitationen auf; diese Phänomene müssen zunächst auf der Ebene der Worterkennung behandelt werden. Es werden Redewendungen benutzt, die in Texten nicht verwendet werden bzw. als nicht angemessen gelten (z.B. die Verbindung von Sätzen mit 'und zwar' oder die Bestätigung durch Wiederholung von Satzfragmenten des Dialogpartners).

Der für die Verfolgung eines Gesprächsverlaufs notwendige Dialogkontext wird in der Ebene 'Dialog' in Bild 1 durch Führung eines *Dialoggedächtnisses* bereitgestellt. Die Struktur des Dialogs ist in einem *Dialogmodell* definiert, das aus zwei Stichproben von 33 über Telefon geführten Auskunftsdialogen mit zusammen 2700 sec Dauer ermittelt wurde. In Bild 2 wird das Dialogmodell und beispielhaft ein Subnetz gezeigt. Das vollständige Modell findet sich in [Mas93]. Danach läßt sich ein Dialog in die im Bild gezeigten *Dialogschritte* gliedern, die als terminales Alphabet einer Dialoggrammatik aufgefaßt werden können.

Der Dialogschritt 'B_INF_FRG' (Informationsfrage des Benutzers) im Modell kann aus der reinen Frage ('wann geht der nächste Zug nach Frankfurt?') bestehen oder auch noch einen Gruß enthalten ('guten Morgen, wann geht der nächste Zug nach Frankfurt?'). Das Teilnetz 'BESTÄTIGUNG/' dient der gegenseitigen Versicherung, daß alles richtig verstanden wurde. Im Teilnetz 'DEFAULT/' wird vom System nach der Bestätigung des als Standard angenommenen Abfahrtsortes 'Nürnberg' gefragt, wenn dieser nicht angegeben wurde. Im Teilnetz 'NACHFRAGE/' wird vom System gegebenenfalls nach anderen nicht geäußerten obligatorischen Elementen eines Auskunftskonzepts auf der Pragmatikebene gefragt. Im Teilnetz 'PRÄZISIERUNG/' wird vom System eine Einengung der Benutzerfrage gefordert, wenn z.B. wegen einer ungenauen Zeitangabe zu viele Züge gefunden wurden.

Vom Dialogmodul werden folgende Aufgaben übernommen:

- Unter den alternativen pragmatischen Interpretationen wird eine für die weitere Verarbeitung ausgewählt.
- Ein Dialoggedächtnis, das die für den Gesprächsverlauf relevanten Bezugsobjekte enthält, wird aufgefrischt.
- Die zu ermittelnden Fakten für die Antwort werden eingeschränkt.

Die Auswahl einer pragmatischen Interpretation erfolgt auf der Basis der Bewertung und unter Berücksichtigung des aktuellen Dialogschritts. Damit ergibt sich das Problem der Abbildung zwischen Äußerung und Dialogschritt-Typ. Wegen des Dialogmodells sind in jedem Zustand des Dialogs nur einige wenige Dialogschritte möglich. Diese werden in mehreren Stufen wie folgt unterschieden:

1. Metakommunikative Merkmale: Diese sind, falls anwendbar, im Lexikon repräsentiert; z.B. weist 'guten Morgen' sofort auf den Dialogschritt 'S_GRUSS' hin.
2. Syntaktische Realisierung: Es wurden einige syntaktische Konstruktionen ergänzt, die bestimmte Sätze (wie 'nein, danke') bereits auf der Syntaxebene erkennen und durch den entsprechenden Dialogschritt (im obigen Falle 'B_ABLEHNUNG') markieren.
3. Semantische Realisierung: Einige metakommunikative Ausdrücke können relativ frei formuliert werden und werden daher durch spezielle Kasusrahmen des Verbs erkannt, wie z.B. in 'ich danke Ihnen vielmals'.
4. Pragmatische Realisierung: Insbesondere echte Informationsfragen führen zur Instantiierung des zugehörigen pragmatischen Konzepts und können daher als 'B_INF_FRG' oder 'B_ERG' markiert werden.
5. Der Satzmodus kann genutzt werden — zur Zeit werden nur Frage- und Aussagesätze unterschieden.

Auch das Dialogwissen wird im semantischen Netz repräsentiert. Bild 3 zeigt die Realisierung einiger Dialogschritte im Netz.

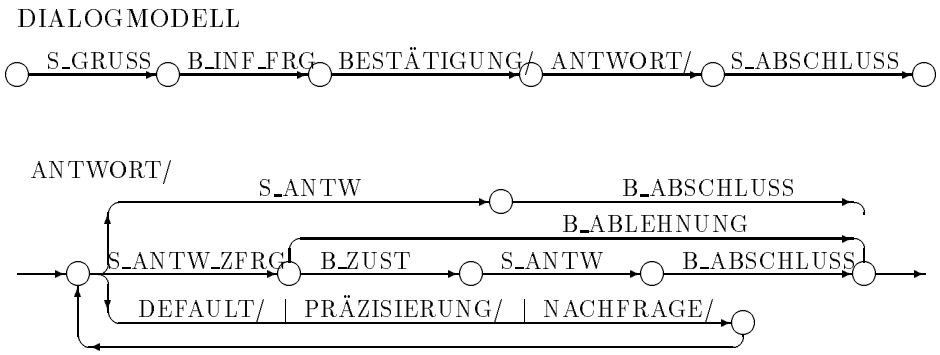

Abbildung 2: Vereinfachte Version des Dialogmodells

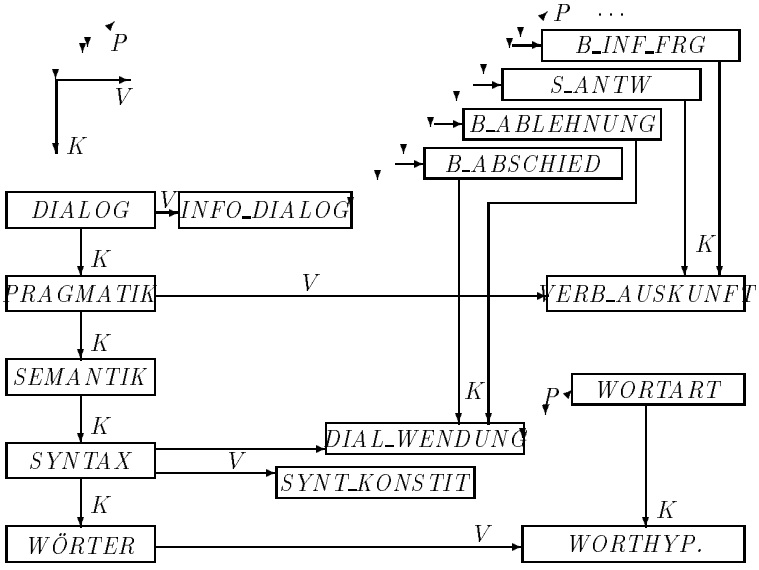

Abbildung 3: Ein *Ausschnitt* aus der Wissensbasis, der u.a. einige Dialogschritt-Typen zeigt, die als Bestandteile des Konzepts *INFO_DIALOG* angeordnet sind. Man erkennt auch deutlich die dreidimensionale Strukturierung der Wissensbasis längs der Kanten der Konkretisierungen K, Spezialisierungen V und Bestandteile P

2.6 Antwort

Die Generierung einer natürlichsprachlichen Antwort und deren Umsetzung in synthetische Sprache stehen nicht im Vordergrund der Arbeiten. Daher werden Antworten hier einfach mit Textschablonen generiert, und für die Sprachsynthese wird ein kommerzielles Gerät verwendet.

Ein vollständig instantiiertes Pragmatikkonzept enthält alle Information, die zur Suche nach einer Zugverbindung erforderlich ist. Als Datenbank zur Ermittlung der gewünschten Verbindungen wird das - auch von der Deutschen Bundesbahn verwendete - HAFAS-System[1] eingesetzt. Die Fakten der Antwort werden als Werte von Variablen in Antwortschablonen eingesetzt. Diese Art der Antwortgenerierung ist einfach, wird aber für diese Anwendung zumindest im Augenblick als hinreichend angesehen.

[1] Die Datenbank wurde entwickelt von: HaCon, Hannoversche Consulting für Verkehrswesen, Transporttechnik und Elektronische Datenverarbeitung GmbH, Hannover.

3 Bewertungen in einem sprachverstehenden System

Für eine effiziente Analyse ist die Bewertung von Zwischenergebnissen erforderlich. Da die Verarbeitung in den beiden Phasen der datengetriebenen Worterkennung und der modellbasierten linguistischen Analyse durchgeführt wird, wird im folgenden auf beide Phasen eingegangen. Bei der linguistischen Analyse ist die Bewertung einer Instanz eines Konzepts von der eines Knotens im Suchbaum zu unterscheiden.

3.1 Bewertung von Worthypothesen

Die Bewertung von Wörtern und Wortketten — im folgenden einheitlich als *Worthypothesen* bezeichnet — ist die Basis der Bewertung in der nachfolgenden linguistischen Analyse. Die Bewertung einer Worthypothese ist die Summe der Bewertung des Teils der Äußerung, der durch die Worthypothese *überdeckt* ist ($\hat{\psi}$), und des Teils, der noch *unüberdeckt* ist ($\hat{\chi}$). Der erste Term basiert auf der aus der Worterkennung ermittelten Güte der Worthypothese G_q. Für den zweiten Term wird ein in den meisten Fällen optimistischer Schätzwert G_r bestimmt (siehe dazu [ST87]).

Eine Komponente des Bewertungsvektors $\boldsymbol{\phi}$ für einen Knoten im Suchbaum ist daher die *Qualität*

$$\hat{\phi}_q = G_q + G_r = \hat{\psi} + \hat{\chi} \qquad (1)$$

Sie genügt in den meisten Fällen den Anforderungen des A^*-Algorithmus.

3.2 Bewertung der linguistischen Analyse

Während der Worthypothetisierung werden die N Hypothesen bestimmt, die gemäß G_q am besten sind. Der erste Schritt der linguistischen Analyse wählt darunter die N' *pragmatisch relevanten* Wörter aus. Die weiteren Schritte der linguistischen Analyse erzeugen daraus linguistisch sinnvolle Wortfolgen, z.B. syntaktische Konstituenten, Kasusrahmen von Verben, Instanzen pragmatischer Konzepte (wie *FAHRPLAN_AUSKUNFT*). Jede solche Wortfolge bildet eine neue Instanz I des zugehörigen Konzepts im Modell, die ihrerseits zusammen mit möglichen Konzeptmodifikationen in einem Suchbaumknoten v untergebracht wird.

In jedem Schritt der Analyse wird der Schätzwert $\hat{\boldsymbol{\phi}}$ der Bewertung des Suchbaumknotens von der aktuellen Bewertung des modifizierten oder instantiierten *Zielkonzepts* C_g abgeleitet. Die Bewertung $\mathbf{G}$ einer Instanz $I(C)$ oder eines modifizierten Konzepts $Q(C)$ ist der Vektor $\mathbf{G} = (G_c, G_q)_t$ mit den Komponenten

- G_c: Kompatibilität der Hypothese mit dem linguistischen Wissen (eine binäre Zahl),
- G_q: Qualität der Wortfolge, die aus nicht notwendig zeitlich benachbarten Wörtern besteht.

Die Bewertung eines Suchbaumknotens v beruht auf der Vorgabe, daß das berechnete Ergebnis mit dem linguistischen Wissen und dem Dialogkontext kompatibel sein und optimal zum Sprachsignal passen soll.

Die Bewertung $\hat{\boldsymbol{\phi}}$ eines Knotens v mit zugeordnetem Zielkonzept C_g und aktueller Modifikation $Q(C_g)$ ist der Vektor

$$\hat{\boldsymbol{\phi}} = (\phi_c, \phi_q, \phi_r, \phi_t)_t \,. \qquad (2)$$

Er wird vom Suchalgorithmus in lexikographischer Ordnung ausgewertet.

Die Komponenten des Bewertungsvektors sind:

- ϕ_c: *Kompatibilität* der Hypothese mit dem linguistischen Wissen

  $$\phi_c = G_c(Q(C_g)) \in \{0,\, 1\} \,,$$

  da die weitere Verarbeitung anderer Hypothesen nicht sinnvoll ist.

- ϕ_q: Vom Sprachsignal abgeleitete *Qualität* (1) der Wortkette, die die Hypothese bildet

  $$\phi_q = G_q(Q(C_g)) + G_r(Q(C_g)) \,,$$

  d.h. unter den kompatiblen Hypothesen wird die am besten zum Signal passende bevorzugt.

- ϕ_r: *Zuverlässigkeit* der Hypothese

  ϕ_r = [Zahl der Rahmen (10ms Dauer) in der längsten Kette zeitlich benachbarter Wörter],

  da unter den Ketten mit hoher Qualität die zuverlässigeren zu bevorzugen sind; es wurde experimentell verifiziert, daß lange Wörter besser erkannt werden als kurze.

- ϕ_t: *Überdeckung* der Äußerung durch die Hypothese

  ϕ_t = [Zahl der durch Worthypothesen überdeckten Sprachrahmen in $Q(C_g)$],

  denn unter den zuverlässigen Hypothesen werden die bevorzugt, die wegen ihrer grösseren Überdeckung vermutlich rascher zu einem Endergebnis führen.

Die Bewertung $\hat{\boldsymbol{\phi}}$ in (2) beruht einerseits auf den theoretisch begründeten Anforderungen des A^*-Algorithmus und wurde zur Effizienzsteigerung andererseits um heuristische Kriterien erweitert. Die Experimente zeigen, daß damit eine erfolgreiche Analyse gesprochener Sprache möglich ist.

4 Ergebnisse und Ausblick

Das in Bild 1 gezeigte System wurde mit dem in Abschnitt 2 beschriebenen linguistischen Wissen realisiert. Die generelle Systemarchitektur des Spracherkennungs- und Dialogsystems wurde in [Nie86] vorgestellt.

Ein Dialogsystem, welches eine vollständige Verarbeitung vom Sprachsignal zur Dialogführung erlaubt wurde realisiert [Kum91, Mas93]. Das System ermöglicht einen gesprochenen Dialog zum Thema Fahrplanauskunft mit einem Benutzer, wobei die Spracherkennung sprecherunabhängig arbeitet und die Systemäußerungen in synthetisierter Sprache ausgegeben werden.

Die Realisierung erfolgte unter UNIX in C ohne Verwendung spezieller Hardware (abgesehen von der Schnittstelle für die Spracheingabe). Die CPU-Zeit für die gesamte Verarbeitung einer Benutzeräußerung bis zur Generierung einer Systemreaktion hängt von der Qualität der gesprochenen Sprache und der Komplexität der Anfrage ab.

Dieses System wurde in einer sprecherunabhängigen Version mit je fünf Dialogen von 2 männlichen und 2 weiblichen Sprechern getestet. Die Spracherkennung arbeitete mit einem Vollformenlexikon mit ca. 1000 Einträgen und verwendete kein Sprachmodell, d.h. die Perplexität betrug ebenfalls ca. 1000. Von den 20 Dialogen konnten 85% mit Erfolg geführt werden. Dabei gab es in 20% der Dialoge mindestens einmal eine Fehlinterpretation, die jedoch über die Dialogführung wieder behoben wurde. Nur in 15% der Dialoge konnte keine Zugauskunft generiert werden.

Mit diesen Arbeiten konnte eine Systemversion realisiert werden, die eine gemischte daten- und modellgetriebene Analysestrategie verwendet. In einem anderen Vorhaben wurde das System um ein Prosodie-Modul ergänzt, welches eine prosodisch basierte Dialogsteuerung ermöglicht [Kom93]. Weitere Arbeiten werden sich auf die Verbesserung der Worterkennung v.a. in bezug auf spontansprächliche Phänomene wie etwa Häsitationen, die Erweiterung des Sprachumfangs und den Ausbau der Dialogfähigkeit konzentrieren.

Literatur

[Abr78] W. Abraham (Hrsg.): *Valence, Semantic Case,*

and Grammatical Relations, Vol. 1, John Benjamins, Amsterdam, 1978.

[All89] J. Allgayer, K. Harbusch, A. Kobsa, C. Reddig, N. Reithinger, D. Schmauks: *XTRA: A Natural Language Access System to Expert Systems*, *Int. Journ. of Man Machine Studies*, Bd. 31, 1989, S. 161–195.

[Bah83] L. Bahl, F. Jelinek, L. Mercer: *A maximum likelihood approach to speech recognition*, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Bd. 5, 1983, S. 179–190.

[Bat93] M. Bates, R. Bobrow, P. Fung, R. Ingria, F. Kubala, J. Makhoul, L. Nguyen, R. Schwartz, D. Stallard: *The BBN/HARC Spoken Language Understanding System*, in *Proc. Int. Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, MN, 1993, S. II–111 – II–114.

[Bri84] A. Brietzmann: *Semantische und pragmatische Analyse im Erlanger Spracherkennungsprojekt*, Arbeitsberichte des IMMD Band 17, Nr. 5, Dissertation, Technische Fakultät, Universität Erlangen-Nürnberg, Erlangen, 1984.

[Bri87] A. Brietzmann: *Stufenweise syntaktische Analyse mit integrierter Bewertung für die kontinuierliche Spracherkennung*, Arbeitsberichte des IMMD Band 20, Nr. 9, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg, Erlangen, 1987.

[Ehr86] U. Ehrlich: *Ein Lexikon für das natürlichsprachliche Dialogsystem EVAR*, Bd. 19 von *Arbeitsberichte des Inst. für Mathematische Maschinen und Datenverarbeitung*, Universität Erlangen-Nürnberg, Erlangen, F. R. of Germany, 1986.

[Ehr90] U. Ehrlich: *Bedeutungsanalyse in einem sprachverstehenden System unter Berücksichtigung pragmatischer Faktoren*, Sprache und Information Bd. 22, Max Niemeyer, Tübingen, 1990.

[Fil68] C. Fillmore: *The Case for Case*, in E. Bach, R. Harms (Hrsg.): *Universals in Linguisitic Theory*, Holt, Rinehardt, and Winston, New York, 1968, S. 1–90.

[Hir89] L. Hirschman, F.-M. Lang, J. Dowding, C. Weir: *Porting PUNDIT to the Resource Management Domain*, in *Speech and Natural Language Workshop*, Philadelphia, 1989, S. 277–282.

[Kom93] R. Kompe, A. Kiessling, T. Kuhn, M. Mast, H. Niemann, E. Nöth, K. Ott, A. Batliner: *Prosody takes over: A Prosodically Guided Dialog System*, Berlin, Germany, 1993, S. 2003–2006.

[Kum91] F. Kummert: *Flexible Steuerung eines sprachverstehenden Systems mit homogener Wissensbasis*, Bd. 12 von *Dissertationen zur Künstlichen Intelligenz*, Infix, Sankt Augustin, 1991.

[Kun91] S. Kunzmann: *Die Worterkennung in einem Dialogsystem für kontinuierlich gesprochene Sprache. Dissertation, Technische Fakultät, Universität Erlangen-Nürnberg*, Bd. 264 von *Linguistische Arbeiten*, Max Niemeyer, Tübingen, 1991.

[Low76] B. Lowerre: *The HARPY Speech Recognition System*, Dissertation, Dept. Comput. Sci., Carnegie-Mellon University, Pittsburgh, PA, 1976.

[Mas93] M. Mast: *Ein Dialogmodul für ein Spracherkennungs- und Dialogsystem*, Bd. 50 von *Dissertationen zur Künstlichen Intelligenz*, Infix, Sankt Augustin, 1993.

[Mas94] M. Mast, F. Kummert, U. Ehrlich, G. Fink, T. Kuhn, H. Niemann, G. Sagerer: *A Speech Understanding and Dialog System with a Homogeneous Linguistic Knowledge Base*, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1994, S. 179–194.

[Nie85] H. Niemann, A. Brietzmann, R. Mühlfeld, P. Regel, G. Schukat: *The Speech Understanding and Dialog System EVAR*, in R. DeMori, C. Y. Suen (Hrsg.): *New Systems and Architectures for Automatic Speech Recognition and Synthesis*, Bd. 16 von *NATO ASI Series F*, Springer, Berlin, Heidelberg, New York, Tokyo, 1985.

[Nie86] H. Niemann, A. Brietzmann, U. Ehrlich, G. Sagerer: *Representation of a continuous speech understanding and dialog system in a homogeneous semantic net architecture*, in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, Tokyo, Japan, 1986, S. 30.6.1–30.6.4.

[Nie88] H. Niemann, A. Brietzmann, U. Ehrlich, S. Posch, P. Regel, G. Sagerer, R. Salzbrunn, G.Schukat-Talamazzini: *A knowledge based speech understanding system*, *Int. Journal of Pattern Recognition and Artificial Intelligence*, Bd. 2, 1988, S. 321–350.

[Nöt91] E. Nöth: *Prosodische Information in der automatischen Spracherke nnung - Berechnung und Anwendung*, Max Niemeyer Verlag, Tübingen, 1991.

[Rab88] L. Rabiner: *Mathematical foundations of hidden Markov models*, in H. Niemann, M. Lang, G. Sagerer (Hrsg.): *Recent Advances in Speech Understanding and Dialog Systems*, Bd. 46 von *NATO ASI Series F*, Springer, Berlin, 1988, S. 183–205.

[Sag90] G. Sagerer: *Automatisches Verstehen gesprochener Sprache*, Bd. 74 von *Reihe Informatik*, BI Wissenschaftsverlag, Mannheim, 1990.

[ST87] E. Schukat-Talamazzini: *Generierung von Worthypothesen in kontinuierlicher Sprache*, Bd. 141 von *Informatik Fachberichte*, Springer, Berlin, 1987.

[ST94] E. Schukat-Talamazzini: *Automatische Spracherkennung*, erscheint bei Vieweg, Wiesbaden, 1994.

[Tes66] L. Tesniere: *Elementes des Syntaxe Structurale*, Klincksieck, Paris, 1966.

[Wah88] W. Wahlster: *Natural Language Systems. Some Research Trends*, Bericht 43, Universitä des Saarlandes, Sonderforschungsbereich Künstliche Intelligenz, Saarbrücken, 1988.

[You89] S. Young, C. Proctor: *The Design and Implementation of Dialogue Control in Voice Operated Database Inquiry Systems*, *Computer Speech & Language*, Bd. 3, Nr. 4, 1989, S. 329–353.