Siegfried Kunzmann, Die Worterkennung in einem Dialogsystem für kontinuierlich gesprochene Sprache, Niemeyer, Tübingen 1991 (Linguistische Arbeiten 264), ISBN 3-484-30264-X, 200 pp., Kart. DM 86,--

Einleitung - Automatische Worterkennung

Für uns alle, die wir eine oder mehrere Sprachen beherrschen, stellt die praktische Durchführung der Worterkennung im Regelfall kein Problem dar: Ein trainierbarer "Mechanismus" - oder, nach modernerer Auffassung, ein "Modul" - in unserem Gehirn sorgt dafür, daß aus dem komplizierten, akustischen (Sprach-)Schallsignal, das vom Ohr wahrgenommen wird, die für die syntaktische und semantische Verarbeitung nötigen Einheiten, Segmente - i.a. die sog. "Wörter" - gefiltert und bestimmt werden.

Daß dieser dafür notwendige "Mechanismus" bzw. Prozeß alles andere als einfach ist, wird klar, wenn man Versuche unternimmt, kontinuierlich gesprochene Sprache durch Computer zu erkennen - d.h. zunächst zu analysieren, dann zu segmentieren und schließlich auch zu verstehen.¹

Im Zusammenhang mit Kunzmanns Buch interessieren uns der erste und zweite Schritt; die hier betrachtete Aufgabenstellung lautet: "Wie gelangt man vom vorliegenden akustischen Sprachsignal zu einer für den Parser geeigneten linearen Repräsentation in Form einer Wortkette?"

"Die Worterkennung ist im Erkennungssystem [von EVAR, s.u.] das Bindeglied zwischen der akustisch-phonetischen Erkennung und den Modulen zur Interpretation von Wortketten. Sie hat ganz generell die Aufgabe, Wörter oder Wortketten im Signal zu finden bzw. mit dem Signal zu vergleichen, wobei ein Lexikon und ein geeignet repräsentiertes Sprachsignal zugrundegelegt wird", Kunzmann (p.15).

Kunzmanns Arbeit ist in das seit etwa einem Jahrzehnt an der Universität Erlangen-Nürnberg laufende Projekt EVAR² integriert, das einen

1. Neben diesem analytischen Ansatz gibt es auch den holistischen Ansatz der sog. Neuronalen Netze, wo viele einfache Subsysteme ("Computer-Neuronen") zu komplexen, lernfähigen und trainierbaren Netzwerken verbunden sind. Kunzmann erwähnt bemerkenswerterweise diese Neuronalen Net-

automatisierten Mensch-Maschine-Dialog zum Ziel hat, wobei folgende wichtige Randbedingungen³ zu berücksichtigen sind:

- (1) Sprecherunabhängigkeit: das System muß in der Lage sein, die Benutzeräußerungen ohne vorangehende Adaptionsphase zu erkennen
- (2) ausreichender Wortschatz (bei EVAR sind zur Zeit ca. 4000 Vollformen im Lexikon)
- (3) Grundlage der Aussprache ist die Standardaussprache nach Duden; es sind zwar (regelhafte) Aussprachevarianten⁴, aber keine Dialekte zugelassen
- (4) Begrenzung des Signals auf Telefonbandbreite (d.h. Berücksichtigung von Frequenzen bis 3,4 kHz)

Das Worterkennungsmodul

Kunzmann beschreibt in seinem Buch ein flexibles Experimentiersystem, welches erlaubt, unterschiedliche Algorithmen, die die Worthypothesen generieren bzw. mit dem Sprachsignal vergleichen, auf ihre Effizienz zu überprüfen. Das Ziel ist, bei möglichst geringem Speicher- und Zeitbedarf - angestrebt wird natürlich eine Erkennung in Echtzeit - eine möglichst hohe Worterkennungsrate zu erreichen. Dabei ist klar, daß der Berechnungsaufwand (mindestens) proportional der Zahl der im *selben* Zeitintervall betrachteten Worthypothesen ist.

Das hier vorgestellte Worterkennungsmodul soll sowohl "datengetrieben" ("von unten", also vom Output des Akustik-Phonetik-Moduls) als auch "erwartungsgesteuert" (Anforderung "von oben", etwa vom Parser oder vom Semantikmodul) arbeiten.

Kunzmann verwendet nun für den Mustervergleich (Worthypothesen vs. Sprachsignal) sog. *elementare Markov-Modelle*, d.h. Wahrscheinlichkeits-

ze in seinem Buch nicht, obwohl auf Laut- und Einzelwortebene damit schon Erfolge in der Spracherkennung vorliegen.

2. Erkennen, Verstehen, Antworten, Rückfragen.

modelle für die Lautmodellierung und die Verkettung von Lauten untereinander. Graphentheoretisch gesehen sind diese Markov-Modelle zyklenfreie, gerichtete und gewichtete Graphen, wobei die Knoten ("Zustände") linear angeordnet sind und der zeitlichen Folge entsprechend über die Kanten erreicht werden: dabei gibt die Kantenmarkierung die Wahrscheinlichkeit der Ausgabe eines hypothetisierten Lautes beim Übergang von einem Knoten zum nächsten an. Diese Wahrscheinlichkeiten (Kantengewichte) müssen in einer Vorlaufphase initialisiert und trainiert werden, um die jeweiligen Lautmodelle möglichst optimal zu approximieren⁵, wobei dieses Training automatisch und/oder händisch erfolgen kann.

Das von einem Akustik-Phonetik-Modul vorklassifizierte Sprachsignal wird nun zunächst segmentiert, d.h. es werden akustisch homogene Bereiche zu Lauten zusammengefaßt. Diese Laut-Segmente werden zu einem sog. Segmentnetz zusammengehängt, worin schließlich ein optimaler Pfad so gesucht wird, daß die Worthypothesen dem Eingabesignal möglichst ähnlich sind.

Bei den besten hier verwendeten Markov-Modellen ergaben sich im Test Worterkennungsraten von über 80 Prozent ab etwa zehn Worthypothesen pro Segment.

In weiteren Kapiteln wird auch auf die Einbettung des Worterkennungsmoduls in die syntaktische (und semantische) Analyse eingegangen.

Neben einer ausführlichen und klaren Zusammenfassung soll der Anhang mit detaillierten Angaben zu Lautklassen, Wortarten, semantischen und pragmatischen Konzepten sowie dem Umfang der in EVAR verwendeten Grammatiken positiv erwähnt werden. Ein Index oder ein Stichwortverzeichnis fehlt leider.

3. Diese ergeben sich zum Teil aus allgemeinen Erwägungen, zum Teil aufgrund der mit EVAR zu realisierenden Aufgabe, eines automatisierten telefonischen Auskunftsystems für Intercity-Zugverbindungen in der Bundesrepublik Deutschland.

4. z.B. "Entstimmlichung von Plosiven im An- und Auslaut".

Schlußbemerkungen

Abschließend seien der gute und übersichtliche Aufbau des Buches positiv hervorgehoben sowie das unserer Meinung nach erfolgreiche Bemühen des Autors, die teilweise komplexe Materie auch dem nichtspezialisierten Leser verständlich zu machen. Zu diesem Zweck werden auch viele aufschlußreiche Abbildungen, Tabellen und Diagramme eingesetzt, wobei hier kritisiert werden muß, daß aus den Legenden nicht immer eine eindeutige Identifizierung der in den Diagrammen dargestellten Kurven möglich ist.

5. Es werden ca. 40 Lautklassen unterschieden.

Klaus Zechner
 Institut für Sprachwissenschaft
 Universität Graz
 A-8010 G r a z