BMBF

Verbmobil

Verbundvorhaben

Die Evaluierung von Spracherkennungssystemen in Deutschland

Erwin Paulus, Michael Lehning

TU Braunschweig

Vm

Report 70

Juli 1995

Juli 1995

Erwin Paulus, Michael Lehning

Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstr. 22
38092 Braunschweig

Tel.: (0531) 391 - 2453
Fax: (0531) 391 - 8218
e-mail: m.lehning@tu-bs.de

Gehört zum Antragsabschnitt: 14.4 Transkription und teilweise Segmentation auf Wortebene

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 N 0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Die Evaluierung von Spracherkennungssystemen in Deutschland

Erwin Paulus, Michael Lehning
Institut für Nachrichtentechnik, TU Braunschweig

1 Einführung

Bei der Bearbeitung von Aufgaben der Mustererkennung lassen sich in mehreren Zusammenhängen Fragen der Beurteilung stellen. Zunächst wäre die Aufgabenstellung selbst zum Gegenstand der Beurteilung zu machen, dann der zur Lösung vorgesehene Ansatz und schließlilch die daraus resultierende Lösung.

Allgemein gesprochen umfaßt die Aufgabenstellung alle Angaben darüber, was unter welchen Einsatzbedingungen anhand welcher "Muster" erkannt werden soll, wobei das Muster meist als Ergebnis eines Zufallsprozesses angesehen wird. Als Beurteilungskriterien wären einerseits das Bayes'sche Risiko und andererseits die Komplexität eines wirklichkeitsgetreuen Modells des Zufallsprozesses nützlich. Das Bayes'sche Risiko stellt das absolute Minimum für die Fehlerrate dar, das bei der gegebenen Aufgabe bestenfalls erreicht werden kann. Es ist insbesondere bei Aufgaben der Spracherkennung leider nicht verläßlich zu schätzen. Auch die Komplexität eines wirklichkeitsgetreuen Modells des mustererzeugenden Prozesses ist in der Praxis meist nicht faßbar. Dies gilt wiederum insbesondere für die zu Zwecken der Spracherkennung betrachteten Muster.

Vergleichsweise einfach erscheint die Beurteilung der resultierenden Lösung. Als Beurteilungskriterien bieten sich die Fehlerwahrscheinlichkeit und verschiedene damit verwandte Größen, wie Verwechslungsmatrix, Rückweisungsrate, Wahrscheinlichkeiten für Einfügungen und Auslassungen und dergleichen an. Zusätzlich kann auch der zur Erkennung notwendige Aufwand für die Beurteilung bedeutsam sein. Unter Verwendung großer Teststichproben lassen sich die Fehlerwahrscheinlichkeit und damit verwandte Größen meist hinreichend genau schätzen. Wichtig ist dabei, daß als Teststichprobe eine Zufallsstichprobe verwendet wird, die von der weiter unten erwähnten Lernstichprobe statistisch unabhängig ist.

Der Lösungsansatz umfaßt im allgemeinen die Annahme eines "Modells" und die Auswahl eines Verfahrens zur Bestimmung der Modellparameter. Das Modell legt die formale Struktur des Erkennungsprozesses fest, Ablauf und Ergebnis dieses Prozesses hängen einerseits von den Modellparametern und andererseits vom jeweils vorliegenden Muster ab. Die Modellparameter werden meist anhand einer Zufallsstichprobe von Mustern, der sogenannten "Lernstichprobe", bestimmt. d.h. sie werden gleichsam aus Beispielen "erlernt". Als Beurteilungskriterien für einen Lösungsansatz kommen in Betracht:

- Das modellbedingte Minimum der Fehlerwahrscheinlichkeit. Die Abweichung dieses bedingten Minimums von dem durch das Bayes'sche Risiko festgelegten absoluten Minimum ist systematischer Natur und könnte dazu dienen, die Angemessenheit des angenommenen Modells an die vorliegende Aufgabe zu beurteilen. Leider läßt sie sich im allgemeinen nicht abschätzen.

- Das verfahrensbedingte Minimum der Fehlerwahrscheinlichkeit. Eine etwaige Abweichung dieses Minimums vom modellbedingten Minimum geht auf das Lernverfahren zurück und ist systematischer Natur. Das verfahrensbedingte Minimum wird im folgenden kurz bedingtes Minimum genannt. Es könnte dazu dienen, für die vorliegende Aufgabe die Angemessenheit des Ansatzes insgesamt (Modell plus Lernverfahren) zu beurteilen. Es ließe sich — wenn auch nur mit großem experimentellen Aufwand — einigermaßen verläßlich bestimmen (siehe unten).

- Eine Art Konvergenzrate des Lernverfahrens, die angibt, wie schnell die erwartete Fehlerwahrscheinlichkeit bei wachsendem Umfang der Lernstichprobe gegen das oben genannte bedingte Minimum strebt. Anstelle einer solchen Konvergenzrate könnte die erwartete Fehlerwahrscheinlichkeit als Funktion des Umfangs der Lernstichprobe betrachtet werden (siehe Bild 1).

  Die Abweichung der Fehlerwahrscheinlichkeit vom bedingten Minimum ist zufallsbedingt. Bei festem Umfang der Lernstichprobe ist die erwartete Fehlerwahrscheinlichkeit der Erwartungswert der Fehlerwahrscheinlichkeit über alle Lernstichproben dieses Umfangs. Sie läßt sich als Mittelwert über mehrere Fehlerwahrscheinlichkeiten schätzen, denen jeweils eine andere unabhängige Lernstichprobe zugrundeliegt. (Die einzelne Fehlerwahrscheinlichkeit kann dabei immer aus ein- und derselben von allen Lernstichproben unabhängigen Teststichprobe geschätzt werden.)

  Wegen des hohen experimentellen Aufwands werden Konvergenzraten bzw. allgemeiner erwartete Fehlerwahrscheinlichkeiten als Funktion des Umfangs der Lernstichprobe bisher kaum betrachtet. Das gilt insbesondere für Ansätze der Spracherkennung.

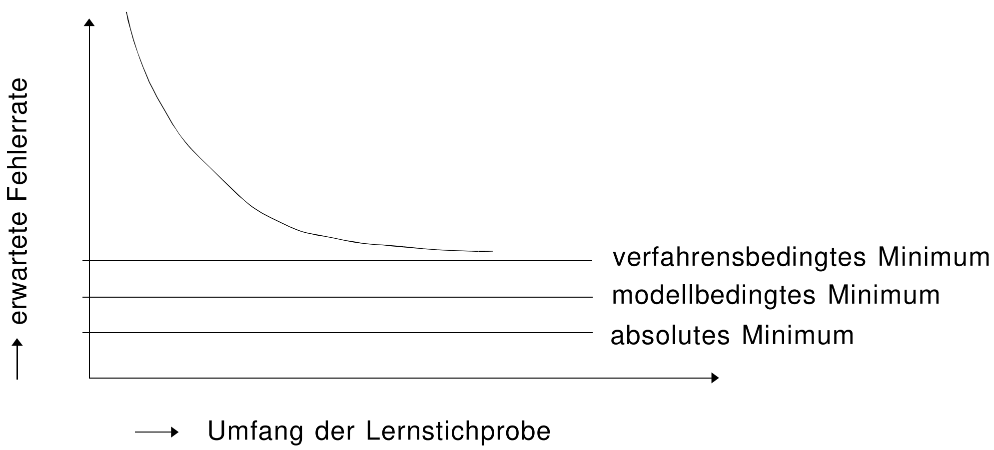

Bild 1: Schematischer Verlauf der erwarteten Fehlerwahrscheinlichkeit als Funktion des Umfangs der Lernstichprobe

Was die vergleichende Beurteilung von Spracherkennungssystemen angeht, werden meist nur die mit "fertigen" Systemen erzielten Fehlerraten verglichen. Dabei gilt es bereits als Fortschritt, daß die verschiedenen Systeme an einer einheitlichen Aufgabe, d.h. aus ein und derselben Teststichprobe, geprüft werden. Wünschenswert — aber nicht immer wirklich sichergestellt — ist es, daß alle Systeme an einer Lernstichprobe gleichen Umfangs "gelernt" haben (siehe die nachfolgenden Abschnitte). Aber selbst unter der zuletzt genannten Bedingung ist die Aussagekraft der Fehlerraten sehr begrenzt und eine darauf aufbauende Rangfolge der Spracherkennungssysteme gilt nur für die betrachtete Aufgabe und da wiederum nur für einen bestimmten Umfang der Lernstichprobe.
Allgemeinere vergleichende Betrachtungen von Spracherkennungssystemen, die auf verschiedenen Ansätzen beruhen, wären auf der Grundlage der erwarteten Fehlerwahrscheinlichkeit (s.o.) als Funktion des Umfangs der Lernstichprobe möglich. Vereinfachend könnte die Lage der Asymptote (bedingtes Minimum, siehe oben) und eine Art Konvergenzrate herangezogen werden. Bild 2 zeigt schematisch zwei fiktive Verläufe der erwarteten Fehlerwahrscheinlichkeit. Daraus ist zu ersehen, daß sich die Rangfolge zweier Lösungsansätze

bei Veränderung des Umfangs der Lernstichprobe umkehren kann. Einer der Ansätze könnte dabei ein vergleichsweise einfaches und wirklichkeitsfernes Modell, der andere ein komplexeres und wirklichkeitsnäheres Modell beinhalten. Das mit dem komplexeren Modell erzielbare bedingte Minimum liegt tiefer als das mit dem einfacheren Modell erzielbare. Der Vorteil des komplexeren Modells kommt im allgemeinen aber erst bei großen Lernstichproben zum Tragen, da sich beim einfacheren Modell (d.h. insbesondere bei einer geringeren Anzahl von Parametern) im allgemeinen eine deutlich raschere Konvergenz bemerkbar macht, so daß schon mit vergleichsweise kleinen Lernstichproben das zugehörige bedingte Minimum erreicht wird.

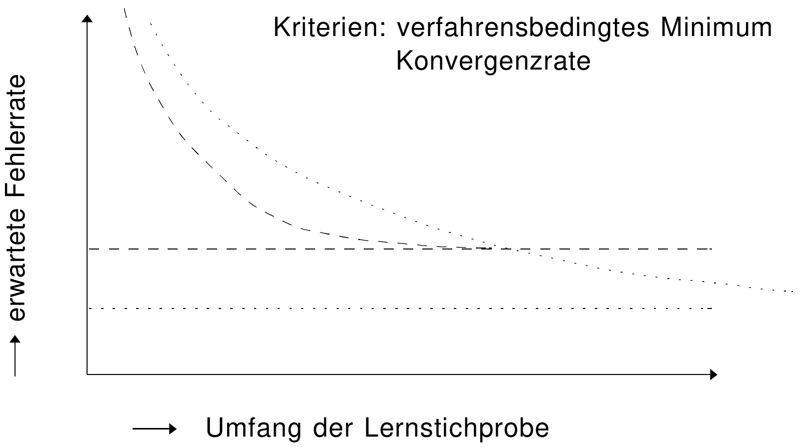

Bild 2: Schematische Verläufe der erwarteten Fehlerwahrscheinlichkeit für zwei verschiedene Lösungsansätze

Noch interessanter wären Beurteilungsverfahren, die auch dann noch einen Vergleich verschiedener Ansätze zulassen, wenn diese nicht an ein und derselben Aufgabe experimentell getestet worden waren. Wenn es gelänge, Aufgaben parametrisch zu beschreiben und experimentelle Testergebnisse so zu interpretieren, daß ein extrapolierbarer Zusammenhang zwischen den Testergebnissen und den Aufgabenparametern hergestellt wird, könnte aus den für eine bestimmte Aufgabe erzielten Ergebnissen auf die Leistungsfähigkeit des Ansatzes auch für andere Aufgaben geschlossen werden. Die in [PAULUS1989] beschriebenen Überlegungen könnten eventuell in diese Richtung weiterverfolgt werden. Das läge aber außerhalb des Rahmens dieses Beitrags, der sich in den folgenden Abschnitten allein auf den praktisch erreichten Stand der Evaluierung von Spracherkennungssystemen konzentriert.

2 Probleme bei der Beurteilung von Spracherkennungssystemen

Eines der größten Probleme bei der Beurteilung von Spracherkennungssystemen liegt in der hohen Anzahl von Faktoren, die einen Einfluß auf die Erkennungsleistung besitzen [LEA1982]. Es erweist sich oft als außerordentlich schwer, den Einfluß eines Faktors auf die Erkennungsleistung zu bestimmen bzw. ein Maß für den Einflußfaktor anzugeben. Die verschiedenen Faktoren lassen sich in folgende Hauptgruppen zusammenfassen:

- **Erkennungsverfahren**: Je nach gestellten Anforderungen an das Erkennungsverfahren ergeben sich deutliche Unterschiede im Schwierigkeitsgrad. Die Verfahren können nach folgendem Schema klassifiziert werden:
  - Erkennungsmodus: Es muß zwischen
    - Einzelworterkennung
    - Verbundworterkennung und
    - Erkennung fließend gesprochener Sprache

    unterschieden werden.

  - Sprecheradaption: Die sprecherunabhängige Erkennung ist wesentlich schwerer als eine sprecherabhängige Erkennung zu realisieren (unter der Annahme gleicher Fehlerraten).
  - Syntax, Semantik, Pragmatik: Die Verarbeitung von Informationen über den Aufbau der Sprache oberhalb der akustischen Ebene kann den Entscheidungsprozeß während der Erkennungsphase vereinfachen, da viele Wortkombinationen von den linguistischen Komponenten des Systems als unzulässig erkannt werden können und somit aus der Weiterverarbeitung herausfallen.

- **Aufnahmeumgebung**: Eine Aufnahme in einem reflexionsarmen Raum, in dem Störgeräusche und Halleffekte auf ein Minimum reduziert sind, liefert nahezu "reine" Sprachaufnahmen. Die Erkennung dieser "reinen" Sprachaufnahmen stellt ein System vor wesentlich geringere Anforderungen als ein "Feldversuch".

- **Benutzer**: Sprechmodus und Verhalten des Sprechers haben einen großen Einfluß auf die Erkennungsleistung. Nicht nur die Sprech- bzw. Klangqualität sind dabei von besonderer Bedeutung sondern auch weitere sprecherbezogene Einflüsse:
  - Vertrautheit mit der Systemumgebung
  - Psychischer Zustand bzw. Motivation des Sprechers (Frustration, Aufregung, Langeweile)
  - Pathologie (z.B. Sprachfehler)
  - Dialekt

- **Übertragungsstrecke**: Bei der Übertragung vom Aufnahmeort bis zum System wird das Sprachsignal durch die Eigenschaften der Übertragungsstrecke verfälscht. Es können gewisse Frequenzbereiche angehoben bzw. abgesenkt werden (z.B. beim Telefon). Außerdem kann es bei der Übertragung des Signals zu einer weiteren Rauschüberlagerung kommen.

- **Dialogführung**: Eine gute und geschickte Dialogführung zwischen dem Benutzer und dem System kann die Akzeptanz durch den Benutzer deutlich verbessern und einen zielgerichteten Dialog vereinfachen.

- **Wortschatz und Schwierigkeit des Wortschatzes**: Ein großer Wortschatz und die Verwendung von akustisch sehr ähnlichen Wörtern (z.B. lachen vs. Rachen) können den Erkennungsprozeß wesentlich erschweren.

3 Schnittstellennormierung für Spracherkennungssysteme

Im Erkennungsprozeß können an geeigneten Verarbeitungspunkten Schnittstellen normiert werden, um die Vergleichbarkeit zwischen unterschiedlichen Systemen zu verbessern. Bei geeigneter Normierung können sogar Module verschiedener Systeme gegeneinander ausgetauscht werden. Auf diese Weise läßt sicht die Leistungsfähigkeit einzelner Module in unterschiedlicher Systemumgebung testen.
Für die Definition von Schnittstellen bieten sich folgende Punkte an:

- Aufnahmebedingungen: Normierung von Mikrophonen, Raumverhältnissen, Audiogeräten, der Abtastrate und der Digitalisierungsauflösung.

- Sprecher: Verwendung von synthetischer Sprache. Dies erlaubt eine hohe Reproduzierbarkeit und ermöglicht eine besonders gut zu kontrollierende Möglichkeit zur Variation sprecherabhängiger Merkmale (z.B. Lage der Formanten, Grundfrequenz und deren Verlauf u.ä.). Eine weitere Möglichkeit stellt der Offline-Test von Spracherkennern dar, in dem jedem zu untersuchenden Spracherkenner die *gleichen* Sprachdaten, die zuvor einmal aufgenommen wurden, angeboten werden.

- Akustische Merkmalsextraktion: Zum Vergleich der akustischen Komponenten eines Spracherkennungssystems können die Merkmale normiert werden, d.h. man gibt jedem System die gleichen Merkmale vor (z.B. Barkskala, Frequenzspektrum, Lineare Prädiktionskoeffizienten).

- Word-Lattice: In der nächsten Ebene könnten die Word-Lattices, die durch die akustische Analyse gewonnen wurden, normiert werden, d.h. jeder nachgeschalteten linguistischen Komponente werden die gleichen Word-Lattices angeboten. In einer Word-Lattice werden alle bewerteten Worthypothesen mit ihren Anfangs- und Endzeitpunkten für den zu testenden Satz ausgegeben.

- Satzebene: Jedem System werden Sätze in orthographischer Notation vorgegeben, die von der Linguistikkomponente ausgewertet werden. Als Ergebnis dieser Analyse wird eine semantische Repräsentation bzw. eine Datenbankanfrage generiert.

4 Auswertung von Erkennungsergebnissen

In der ersten Stufe können die Erkennungsergebnisse numerisch evaluiert werden. Dabei ist die Angabe einer Erkennungsrate nicht besonders aussagekräftig, da sie keine Informationen über die Art der auftretenen Fehler bzw. über die Fehlerverteilung liefert.

4.1 Beurteilung der Erkennungsergebnisse bei Isoliertworterkennern

4.1.1 Bestimmung der Fehlertypen

Betrachtet man die Spracherkennung als eine Aufgabe aus dem Bereich der Mustererkennung, bei der zu einem Muster (Testäußerung) ein Referenzmuster gesucht wird, kann grundsätzlich zwischen den folgenden Fehlertypen unterschieden werden [MANGOLD1990], [SIMPSON1987]:

- Korrekte Erkennung: Das Muster wurde dem entsprechenden Referenzmuster zugeordnet.

- Falsche Erkennung: Das Sprachmuster ist zwar zugelassen, wurde aber dem falschen Referenzmuster zugeordnet (Ersetzung).

- Falschakzeptanzen: Ein nicht gültiges Sprachmuster (z.B. ein dem System unbekanntes Wort) wird einem gültigen Referenzmuster zugeordnet.

- Falsche Rückweisung: Ein zulässiges Sprachmuster wird als unzulässig zurückgewiesen.

- Korrekte Rückweisung: Ein unzulässiges Sprachmuster wird als unzulässig zurückgewiesen.

4.1.2 Ermittlung von Verwechslungsmatrizen

Noch detailliertere Informationen erhält man aus der Bestimmung von Verwechslungsmatrizen. Wie in [DRESCHLER1986] vorgeschlagen wird, kann man aus den Verwechslungsmatrizen Ähnlichkeitsmatrizen erzeugen, wobei ein hohe Ähnlichkeit mit einer hohen Wahrscheinlickkeit für eine Vertauschung einhergeht.

4.2 Beurteilung der Erkennungsergebnisse bei Verbundworterkennern

Für die Ermittlung der Erkennungsleistung von Verbundworterkennern können verschiedene Parameter herangezogen werden [PALLETT1985]:

- Die prozentuale Rate der komplett richtig erkannten Sätze im Verhältnis zu allen getesteten Sätzen

- Prozentrate der richtig erkannten Wörter zu allen Wörtern

- Angabe der Raten für Einfügungen, Löschungen und Ersetzungen von Wörtern innerhalb eines Satzes. Um diese Raten zu bestimmen, müssen die erkannten Sätze zu den Referenzsätzen ausgerichtet werden (String Alignment). Nach der erfolgten Ausrichtung können die Fehlerraten ermittelt werden.

Für die Beurteilung von Verbundworterkennern hat sich die wortweise Bewertung auf Textebene etabliert, so wie sie im amerikanischen DARPA-Projekt durchgeführt wird [PALLETT1985]. Bei diesem Vergleich wird der Referenzsatz mit einem vom System gelieferten Hypothesesatz verglichen. Vor dem eigentlichen Vergleich müssen die Sätze noch gegeneinander ausgerichtet werden [OKUDA1976].

Eine einfache Links-nach-Rechts-Zuordnung ist dabei sehr nachteilhaft, da beim Auslassen bzw. Einfügen von nur einem Wort alle folgenden Wörter als Fehler gezählt werden. Beispiel:

Referenz:	Dies ist ein Test für ein System
Hypothese:	Dies ist Test für ein System

Bei einer Links-nach-Rechts-Ausrichtung ergibt sich dann folgende Zuordnung:

Referenz:	Dies	ist	ein	Test	für	ein	System
Hypothese:	Dies	ist	Test	für	ein	System	
Auswertung:	kor.	kor.	Sub.	Sub.	Sub.	Sub.	Del.

Erläuterung: kor.: korrekt Sub.: Ersetzung Del.: Löschen eines Wortes

Werden die Wörter dagegen mittels der Methode der Dynamischen Programmierung gegeneinander ausgerichtet, so ergibt sich eine wesentlich plausiblere Ausrichung. Die Kosten der Dynamischen Programmierung ergeben sich aus dem Auslassen, Einfügen oder Ersetzen eines Wortes. Bei dem gegebenen Beispiel ergibt sich dann folgende Ausrichtung:

Referenz:	Dies	ist	ein	Test	für	ein	System
Hypothese:	Dies	ist	***	Test	für	ein	System
Auswertung:	kor.	kor.	Del.	kor.	kor.	kor.	kor.

Erläuterung: kor.: korrekt Sub.: Ersetzung Del.: Löschen eines Wortes

4.3 Beurteilung von Worthypothesennetzen (Word-Lattices)

Bei der Beurteilung von Worthypothesenetzen kann prinzipiell in der gleichen Weise wie bei der Beurteilung von Worthypothesenketten vorgegangen werden. Doch ist die Ermittlung von Einfügungen, Auslassungen und Verwechslungen bei Worthypothesennetzen nicht eindeutig.
Um die Eindeutigkeit zu gewährleisten, wird aus dem Worthypothesennetz der Satz ermittelt, der mit dem Referenzsatz die größtmögliche Übereinstimmung liefert. Es wird also nach der Satzhypothese gesucht, die die geringste Gesamtanzahl von Einfügungen, Auslassungen und Vertauschungen liefert. Eingefügte Pausen werden nicht als Einfügungsfehler gezählt.
Dies bedeutet unter anderem, daß sich für einen Satz eine Worterkennungsrate von 100% ergibt, wenn die gesprochene Wortkette im Graphen enthalten ist. Je größer das Netz ist, desto kleiner ist die zu erwartende Fehlerrate. Aus diesem Grunde ist es notwendig, die Hypothesenanzahl auf die Äußerungslänge zu normieren. Es muß also eine *Hypothesendichte* berechnet werden. Zum Vergleich von Algorithmen, die Worthypothesengraphen erzeugen, wird sinnvollerweise die Worterkennungsrate über die Worthypothesendichte als Vergleichsmaß herangezogen (d.h. Anzahl der Worthypothesen für alle Wörter in diesem Satz im Verhältnis zu der Anzahl der (Referenz-)wörter im Satz).

5 Durchgeführte Spracherkennungstests in Deutschland

In Deutschland wurde zum ersten Mal im Rahmen des BMFT-Verbundprojekts ASL bzw. des Nachfolgeprojektes VERBMOBIL eine Evaluierung der signalnahen Spracherkennung durchgeführt. Beginnend mit dem Winter 1992/1993 wurden bisher in halbjährlichem Abstand drei Erkennungstests durchgeführt. An diesen Tests beteiligten sich vier Forschungsinstitutionen. Für den Test und das Training wurde das zu verwendende Vokabular und das Trainings- bzw. Testmaterial vorgegeben.
Als Testmaterial wurden 500 Sätze aus der Domäne Intercity-Auskunft ausgewählt. Die Ergebnisse der Erkennungsläufe wurden zentral gesammelt und ausgewertet.
Von den Institutionen wurden die Erkennungsergebnisse entweder als "beste" Sätze oder als Worthypothesennetze unterschiedlicher Worthypothesendichte angeliefert. Bei den Wort-

hypothesenetzen waren zur besseren Vergleichbarkeit Worthypothesendichten von 5, 10, 20 oder 40 Hypothesen pro Wort anzustreben.
Um die Verbesserung der Erkennungsleistung beim Einsatz eines (statistischen) Language Model zu beurteilen, konnte für die Erkennung ein Bigrammodell benutzt werden.

5.1 Bewertung der Erkennungsergebnisse

Für die angelieferten Erkennungsergebnisse wurden folgende Fehlerraten ermittelt:

- Satzfehlerrate
- Wortfehlerrate

Die Ermittlung der Fehlerraten wurde wie folgt durchgeführt:

1. Bei den Worthypothesennetzen wurde der Hypothesesatz mit der größtmöglichen Übereinstimmung zum Referenzsatz ermittelt. Als Maß für die Übereinstimmung wurde die Levenstheindistanz [OKUDA1976] gewählt, bei der die Auslassung, Einfügung und Vertauschung von Wörtern als gleichwertige Fehler eingestuft werden.
2. Die Referenz- und Hypothesesätze wurden mit der Dynamischen Programmierung (DP) gegeneinander ausgerichtet. Als Distanzmaß wurde wiederum das Levensthein-Maß (s.o.) gewählt.
3. Die Ergebnisse des Vergleichs wurden in maschinenlesbarer und dokumentarischer Form gespeichert (siehe Anhang).

Die Ergebnisse der Evaluierung und die für die Evaluierung erstellte Software ist frei verfügbar und ermöglicht somit auch die Überprüfung der Testergebnisse vor Ort.

5.2 Ausblick

Für die Zukunft ist weiterhin eine Evaluierung in halbjährlichen Abständen vorgesehen. Der Test wird aber dann mit spontansprachlichen Daten durchgeführt, die zur Zeit im Rahmen des Verbmobil-Projektes gesammelt werden.
Es kann aber bereits jetzt festgestellt werden, daß die zentrale Auswertung und Festlegung von Test- und Trainingsmaterial die Vergleichbarkeit der Erkenner für die deutschen Sprache ermöglicht hat und eine Einordnung der Erkennungsleistung der verschiedenen Systemen im nationalen Vergleich erlaubt.

6 Signifikanztests

6.1 Berechnung von Konfidenzintervallen

Bei der Angabe von Erkennungsraten handelt es sich in Wirklichkeit nur um eine Schätzung der Erkennungsrate, da für eine genaue Angabe theoretisch unendlich viele Versuche durchgeführt werden müßten.
Es ist wesentlich sinnvoller, einen Toleranzbereich für die Erkennungswahrscheinlichkeit anzugeben, innerhalb dessen die tatsächliche Erkennungsrate mit einer relativ hohen statistischen Sicherheit liegt (typisch 95% oder 99%).
Bei den vorliegenden Erkennungsaufgaben kann davon ausgegangen werden, daß der Erkennungsprozeß durch einen Bernoulli-Prozeß (Binomialverteilung) modelliert werden kann.
Den Schätzwert für die Erkennungsrate erhält man aus der Beziehung

$$\hat{p} = \frac{k}{n}$$

mit k = Anzahl der richtig erkannten Muster und n = Anzahl aller Muster.
Für genügend große n kann ein Vertrauensintervall $[p_u, p_o]$ aus der Lösung folgender quadratischer Gleichung ermittelt werden [KREYSZIG1975]:

$$(n + c^2)p^2 - (2k + c^2)p + \frac{k^2}{n} = 0$$

Der Wert c ergibt sich aus der Wahl der Konfidenzzahl $\gamma = 1 - \alpha$ entsprechend folgender Tabelle:

Tabelle zur Berechnung der Konfidenzintervalle

Signifikanzschwelle	γ	α	c
95%	0.95	0.05	1.960
99%	0.99	0.01	2.576
99.9%	0.999	0.001	3.291

Als Lösung für p_u und p_o ergeben sich:

$$p_u = \frac{2k+c^2}{2(n+c^2)} - \sqrt{\left[\frac{(2k+c^2)}{2(n+c^2)}\right]^2 - \frac{k^2}{n(n+c^2)}} \tag{1}$$

$$p_o = \frac{2k+c^2}{2(n+c^2)} + \sqrt{\left[\frac{(2k+c^2)}{2(n+c^2)}\right]^2 - \frac{k^2}{n(n+c^2)}} \tag{2}$$

Die Konfidenzzahl γ gibt die statistische Sicherheit an, mit der die tatsächliche Wahrscheinlichkeit p innerhalb des Intervalls $[p_u, p_o]$ liegt.
Zum Beispiel liegt bei einem $\gamma = 0.99$ in nur einem von hundert Fällen die tatsächliche Erkennungsrate p außerhalb des Konfidenzintervalls $[p_u, p_o]$.

6.2 Mc Nemar's Test

Der McNemar's Test wird benutzt, wenn zwei Systeme am gleichen Testmaterial getestet wurden. Er untersucht, ob zwischen den beiden Erkennern ein signifikanter Unterschied in der Erkennungsleistung besteht. Für jedes zu klassifizierende Testmuster j wird das Ergebnis des einen Erkenners mit dem Ergebnis des anderen Erkenners verglichen. Aus der Auswertung dieser Vergleiche wird die Signifikanz im Unterschied der Erkennungsleistungen berechnet [SACHS1974]. Die Berechnung erfolgt nach dem folgenden Verfahren:

Die Zufallsvariable I_{ij} sei wie folgt definiert:

$$I_{ij} = \begin{cases} 0 & \text{wenn das System } i \text{ das Muster } j \text{ richtig klassifiziert} \\ 1 & \text{sonst} \end{cases}$$

Für die Zufallsvariable $Q_j = I_{1j} - I_{2j}$ gilt dann:

$$Q_j = \begin{cases} -1 & \text{Wenn das System 1 das Muster } j \text{ falsch und} \\ & \text{das System 2 das Muster } j \text{ richtig klassifiziert} \\ 0 & \text{Wenn beide Systeme das Muster } j \text{ entweder richtig} \\ & \text{oder falsch klassifizieren} \\ 1 & \text{Wenn das System 1 das Muster } j \text{ richtig und} \\ & \text{das System 2 das Muster } j \text{ falsch klassifiziert} \end{cases}$$

Unter der Hypothese H_0, daß die beiden Systeme die gleiche Erkennungsrate besitzen, muß gelten:

$$P(Q_j = -1) = P(Q_j = 1) = \frac{1}{2}$$

Aus $Q_j = 0$ kann man keine Informationen für den Unterschied in der relativen Erkennungsleistung der verschiedenen Systeme gewinnen.
Desweiteren seien folgende Zufallsvariablen definiert:

$$\begin{aligned} c_1 &= \text{Anzahl von Mustern mit } Q_j = -1 \\ c_2 &= \text{Anzahl von Mustern mit } Q_j = 1 \\ c &= c_1 + c_2 \end{aligned}$$

c_2 ist binomial verteilt mit den Parametern c und $\frac{1}{2}$. Für ein Experiment, in dem $c_2 = C_2$ und $c = C$ gilt, wird der Signifikanztest nach der Berechnung der Hilfsgröße $p_{McNemar}$ durch den Vergleich von $p_{McNemar}$ mit der Signifikanzschwelle α bestimmt:

$$p_{McNemar} = P(c_2 \geq C_2)|H_0) = \sum_{k=C_2}^{C} \binom{C}{k} \left(\frac{1}{2}\right)^C$$

$$\alpha < p_{McNemar} < (1 - \alpha)$$

Liegt $p_{McNemar}$ außerhalb des angegebenen Intervalls, kann davon ausgegangen werden, daß zwischen den beiden Erkennern signifikante Unterschiede in der Erkennungsleistung bestehen.
Die Wahl von α ist abhängig vom gewählten Signifikanzniveau (siehe Tabelle im vorherigen Abschnitt).

Literatur

[DRESCHLER1986] W. A. Dreschler:*Phonetic Confusion in Quiet and Noise for the Hearing Impaired.* Audiology 25, pp.19-28 (1986)

[KREYSZIG1975] E. Kreyszig:*Statistische Methoden und ihre Anwendung.* 5. Auflage, 1975, Vandenhoeck u. Ruprecht, Göttingen

[LEA1982] W. A. Lea:*What causes speech recognizers to make mistakes.* IEEE Proc. ICASSP, Paris, May 3-5, 1982, Vol.3, 2030-2033

[LINDBERG1991] B. Lindberg, O. Andersen, R.K. Jorgensen, S.W. Danielsen:*Esprit Project 2589 (SAM) - Muliti-lingual Speech Input/Output Assessment, Methodology and Standardisation.* Continuous Speech Recognition. SAM-Doc. No. SAM-IES-062

[MANGOLD1990] H. Mangold:*Die Beurteilung der Leistungsfähigkeit von Spracherkennungssystemen.* Tagungsband Elektronische Sprachsignalverarbeitung, Berlin, Sep. 1990, K. Fellbaum (Hrsg.), 50-61

[OKUDA1976] T.Okuda, E.Tanaka, T.Kasai:*A Method for the Cor- rection of Garbled Words Based on the Levenshtein ein Metric.* IEEE Trans. Comp. 25/2, Feb.1976, 172-178

[PALLETT1985] D. S. Pallett:*Automatic Speech Recognisers (Recognition) Performance Assessment.* National Bureau of Standards Special Publication, Gaithersburg, MD 20895, National Bureau of Standards, Washington, USA

[PAULUS1989] E. Paulus, J. Mudler:*Grundzüge eines sprachkundigen Systems zur automatischen Spracherkennung.* Informationstechnik it, Vol. 31, No.5, 1989, 358-364

[SACHS1974] L. Sachs:*Angewandte Statistik - Planung und Auswertung - Methoden und Modelle.* 4. Auflage, Springer-Verlag Berlin Heidenberg New York, ISBN 3-540-06443-5

[SIMPSON1987] C. A. Simpson:*The phonetic Discrimination Test for Speech Recognizers:Part I&II.* Speech Technology, Vol.3, Nos. 4&5, 48-53, March/April & Oct./Nov. 1987, New York

Dieser Report erschien als Beitrag im Tagungsband:
Fortschritte der Akustik, DAGA 94- Dresden, Teil A, 147 - 156