Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil
Verbundvorhaben

Evaluierung von signalnahen Spracherkennungssystemen für deutsche Spontansprache

Michael Lehning

TU Braunschweig

Vm

Report 161
13. August 1996

13. August 1996

Michael Lehning

Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstr. 22
38092 Braunschweig

Tel.: (0531) 391 - 2476
Fax: (0531) 391 - 8218
e-mail: `lehning@ifn.ing.tu-bs.de`

Gehört zum Antragsabschnitt: 14.4 Evaluierung

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.

Inhaltsverzeichnis

1 Einleitung

Im Rahmen des BMBF-Verbundprojektes Verbmobil werden von verschiedenen Institutionen in Deutschland Spracherkennungssysteme für deutsche Spontansprache entwickelt bzw. bestehende Spracherkennungssysteme für die Anwendung auf deutsche Spontansprache optimiert.

Um den Fortschritt bei der Entwicklung der Systeme zu beurteilen und deren Leistungsfähigkeit zu vergleichen, wird vom Institut für Nachrichtentechnik der Technischen Universität Braunschweig einmal jährlich ein Vergleich der von der beteiligten Institutionen entwickelten Spracherkennungssysteme durchgeführt.

Für den Vergleich werden zuvor der zu verwendende Erkennungswortschatz, das Trainingsmaterial und das Auswertungsverfahren festgelegt.

Der letzte Test fand im Juni 1995 statt. An dem Test beteiligten sich die folgenden fünf Institutionen (in alphabetischer Reihenfolge):

Daimler Benz AG, Forschungszentrum Ulm, Abt. Sprachverstehen, F3M/S
Universität Bielefeld, Technische Fakultät — Angewandte Informatik —
Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5)
Universität Hamburg, FB Informatik, Arbeitsbereich: "Natürlichsprachliche Systeme"
Universität Karlsruhe, Lehrstuhl Prof. Waibel

Bei der Vorstellung der erreichten Erkennungsergebnisse werden die Institutionsangaben anonymisiert.

2 Vorgehen

2.1 Festlegung der Trainingsdaten

Für den Durchlauf im Sommer 1995 standen als Trainingsmaterial 3 CD-ROMs mit Sprachdaten zur Verfügung. Die 3 CD-ROMs entsprechen ca. 10 Stunden Sprache. Zur Initialschätzung der Modelle durften Erkennungssysteme benutzt werden, die an anderem Trainingsmaterial trainiert wurden. Zu dem Trainingsmaterial existieren Verschriftungen und Aussprachewörter. Auf Grundlage dieser Daten wurden die Systeme trainiert. Die Daten wurden an vier Datensammelorten bei universitären Einrichtigungen aufgenommen: Bonn, Kiel, München und Karlsruhe.

2.2 Erkennungswortschatz

Der Erkennungswortschatz umfaßte 3307 Worteinträge. Der Wortschatz deckte alle im Trainingsmaterial vorkommenden Worteinheiten ab. Apostrophierte Worteinheiten wie em 'ne, 'nem' wurden nicht zugelassen.

2.3 Festlegung der Testdaten

Die Testdaten wurden aus einer vierten CD-ROM nach dem Zufallsprinzip ausgewählt und in zwei Testsets unterteilt. *Testset S* enthielt nur Äußerungen mit einer Äußerungslänge kürzer als 15 Sekunden. Insgesamt waren dies 265 Äußerungen. *Testset L* enthielt 66 Äußerungen, die alle länger als 15 Sek. waren.

Beide Testsets umfaßten jeweils ca. 23 Minuten Sprache.

2.4 Sprachmodell

Für die Erkennung konnten verschiedene Sprachmodelle verwendet werden:

- Kein Sprachmodell, d.h. jedes Wort war zu jedem Zeitpunkt gleichwahrscheinlich
- Ein einheitliches Bigramm und Trigramm, die vom Projektpartner Philips, Forschungszentrum Aachen, zur Verfügung gestellt wurden. Die N-Gramme wurden an den Verschriftungen der CD-ROM 1 bis 3 trainiert. Für den Zugriff auf die N-Gramme wurde eine Softwareschnittstelle in der Programmiersprache *C* geschaffen, die die Projektpartner in ihre Programme einbinden konnten.
- Von den teilnehmenden Institutionen selbst erstellte Sprachmodelle

Zusammenfassend läßt sich somit die Evaluierung der signalnahen Spracherkennung in verschiedene Schritte gliedern:

1. Festlegung von Trainingsdaten, Kreuzvalidierungsstichprobe und Wortliste
2. Training der Spracherkennungssysteme mit dem festgelegtem Trainingsmaterial, Adaptierung der Parameter an der Kreuzvalidierungsstichprobe
3. Festlegung der Testdaten
4. Erkennung auf den Testdaten
5. Abgabe der Erkennungsergebnisse
6. Auswertung der Erkennungsergebnisse

3 Auswertung der Erkennungsergebnisse

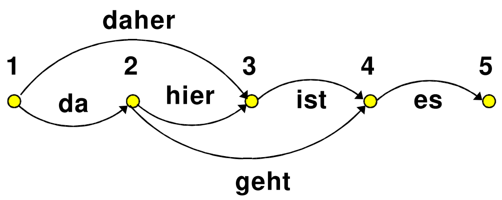
Die Erkennungsergebnisse konnten auf zwei verschiedene Arten abgegeben werden:

- Abgabe des wahrscheinlichsten Satzes

  Als Erkennungsergebnis wird nur der wahrscheinlichste Satz abgegeben.

- Abgabe von Wortgittern

  Es wird ein Wörtergitter als Erkennungsergebnis abgeliefert. Ein Wörtergitter ist ein azyklischer Graph G, der L Kanten besitzt. Jede Kante K_l entspricht einer Worthypothese WH_l und ist gekennzeichnet durch einen Anfangs- und einen Endknoten KA_l und KE_l. Es gibt einen ausgezeichneten Anfangsknoten der Äußerung KAS und einen ausgezeichneten Endknoten der gesamten Äußerung KES. Ein Beispiel für einen Wortgraph mit der korrespondierenden Kantenliste zeigt Bild 1:



Kante (WH)	<von> (KA)	<bis> (KE)
da	1	2
daher	1	3
hier	2	3
geht	2	4
ist	3	4
es	4	5

Startknoten KAS = 1 **Endknoten KES = 5**

Bild 1: Wortgraph mit korrespondierender Kantenliste

Der wahrscheinlichste Satz mit S Wörtern $(W_1 \ldots W_{N_W})$ kann in einen Wortgraph überführt werden, indem folgendermaßen vorgegangen wird:

- Erzeuge N_W Kanten $(K_1, \ldots, K_{N_W})$
- Besetze jede dieser Kanten K_l mit folgenden Werten:
  - Startknoten: $KA_l = l$
  - Endknoten: $KE_l = l + 1$

– Worthypothese: $WH_l = W_l$

Aus jedem Wortgitter wird dann der Pfad ermittelt, der zum Referenzsatz die geringste (gewichtete) Levenshtein-Distanz [OKUDA1976] zum Referenzsatz aufweist. Bei der Fehlerbewertung können dabei für Wortauslassungen (Deletion), -einfügungen (Insertion) und -vertauschungen (Substitution) unterschiedliche Bewertungsmaße ($C_{DEL}, C_{INS}, C_{SUB}$) eingeführt werden. Bei der Evaluierung wurden alle Bewertungsmaße auf 1.0 gesetzt.

Der Pfad, mit den geringsten Kosten, wird mit dem Verfahren der Dynamischen Programmierung nach folgendem Vorgehen ermittelt:

1. Gegeben ist der Referenzsatz mit N_R Referenzwörtern ($RW_1 \ldots RW_{N_R}$ und ein Wortgitter mit L Kanten. Alle vollständigen Pfade beginnen im Knoten $KAS = 1$ und enden im Knoten KES

2. Erzeuge eine Kostenmatrix $C(r,k)$ mit $N_R + 1$ Zeilen ($r = 0, \ldots, N_r$) und KES Spalten ($k = 1 \ldots KES$).

3. Setze $C(r,1) = r * C_{DEL}$

4. Setze $C(r,k) = \infty$ für $k \geq 2$ und $r = 0 \ldots N_r$

5. Für $k = 2 \ldots KES$ prüfe, welche Kanten im Knoten K enden. Für jede Kante, die im Knoten k endet und im Knoten k_s beginnt, berechne $CNEW = C_{ins} + C(0,k_s)$. Falls $CNEW < C(0,k)$ setze $C(0,k) = CNEW$

6. Für $r = 1 \ldots N_r$ durchlaufe die folgende Schleife:

   (a) Für $k = 2 \ldots KES$ Ermittele für alle Kanten, die im Knoten k enden, jeweils den Startknoten k_s, das der jeweiligen Kante zugeordnete Wort W und das Minimum der folgenden Kostenfunktion:
   Wenn $W = RW_r$, dann $CNEW = min(C(r-1,k_s), C(r,ks)+C_{ins}, C(r-1,k) + C_{del})$
   Wenn $W \neq RW_r$, dann $CNEW = min(C(r-1,k_s) + C_{sub}, C(r,ks) + C_{ins}, C(r-1,k) + C_{del})$

   (b) Falls $CNEW < C(r,k)$, dann setze $C(r,k) = CNEW$

Aus $C(N_r, K)$ können die Gesamtkosten ermittelt werden. Wenn man sich für jede Position in der Matrix $C(r,k)$ die entsprechende Vorgängerposition und den Fehlertyp (siehe Bestimmung des Minimums) merkt, kann durch Backtracking die Ausrichtung zwischen Referenzsatz und dem Pfad, der die geringste (gewichtete) Levenshtein-Distanz zum Referenzsatz aufweist, ermittelt werden [PAULUS1994].

Aus dem Ergebnis der Ausrichtung werden die Wortfehlerraten nach folgender Tabelle bestimmt:

Rate	Berechnungsvorschrift
Rate der korrekt erkannten Wörter	#korrekt erkannte Wörter/ #Referenzwörter
Ersetzungsrate:	#ersetzte Wörter/#Referenzwörter
Einfügungsrate:	#eingefügte Wörter/#Referenzwörter
Auslassungsrate:	#ausgelassene Wörter/#Referenzwörter
Wortfehlerrate:	Ersetzungsrate + Einfügungsrate + Auslassungsrate
Wortakkuratheitsrate:	1 − Wortfehlerrate

Es ist einsichtig, daß bei zunehmender Komplexität des Wortgraphen die Wahrscheinlichkeit steigt, daß der korrekte Satz im Wortgraphen enthalten ist. Aus diesem Grund ist es nötig, zu jedem Wortgraphen nicht nur die Fehleranzahl, sondern auch eine Komplexitätabschätzung zu liefern.

Für die Evaluierung im Verbundprojekt Verbmobil wurden zu diesem Zweck die *Worthypothesendichte* und die *durchschnittliche Verzweigungsdichte* ermittelt:

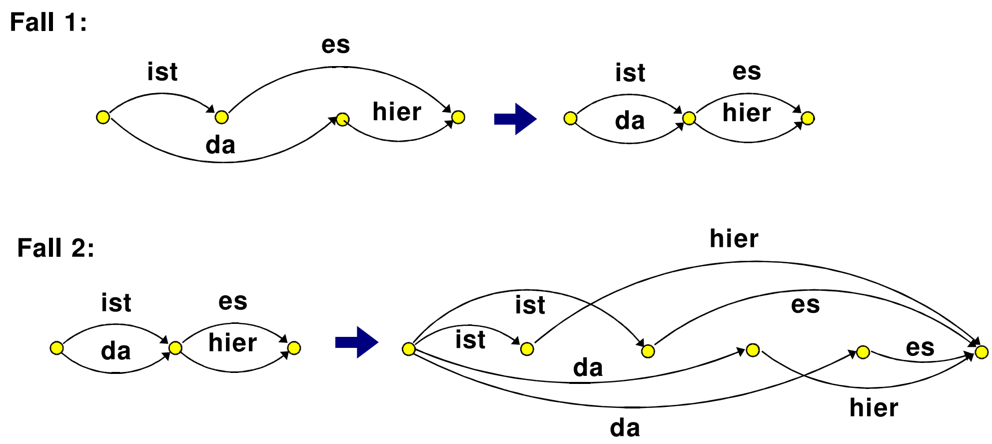
- Worthypothesendichte: Durchschnittliche Anzahl von Hypothesen pro Referenzwort.

- Durchschnittliche Verzweigungsdichte: Durchschnittliche Anzahl von Kanten, die aus einen Knoten heraus- bzw. hineinlaufen

Aus keinem dieser beide Maße kann aber direkt auf die Komplexität des Graphen geschlossen werden. Außerdem ist es möglich, die Komplexitätsmaße bei entsprechender Nachverarbeitung des Graphen zu umgehen. Im ersten Fall kann man versuchen, die Verzweigungsdichte pro Knoten zu erhöhen und somit die Anzahl der möglichen Pfade zu vergrößern (siehe Bild). Im zweiten Fall kann man die Knoten "aufbrechen", indem man Kanten vervielfacht und neue Knoten einfügt (siehe Bild).

Als Ausweg bieten sich mehrere Alternativen an: 1. Berechnung der Gesamtanzahl aller mögliche Pfade durch den Graphen, 2. ein heuristische Maß aus der Kombination der Worthypothesendichte und durchschnittlicher Verzweigungsdichte und 3. eine zufallsgesteuerte Suche im Graphen mit Ermittlung der Verzweigungsdichte bzw. Fehlerraten bei den so ermittelten zufälligen Pfaden (Monte-Carlo-

Methode, augenblicklich noch Gegenstand der Forschung)



Probleme bei der Komplexitätsabschätzung von Wortgraphen

4 Ergebnisse Evaluierung 1995

Die folgenden Tabellen zeigen die Erkennungsergebnisse der verschiedenen Spracherkennungsergebnisse auf den wahrscheinlichsten Sätzen ("first best"). Bei den Wörtergittern liegen die Erkennungsraten naturgemäß noch höher, da ja der wahrscheinlichste Satz stets einen möglichen Pfad in dem zugehörigen Wortgraphen darstellt. Man erkennt, daß der Einsatz eines statistischen Sprachmodells bei allen Institutionen eine deutliche Leistungssteigerung in der Erkennung bewirkt.

Allgemein kann festgestellt werden, daß die Erkennungsraten gegebenüber dem Evaluierungslauf von 1994 um bis zu 20% (absolut) gestiegen sind. Dies entspricht doch einer deutlichen Leistungssteigerung innerhalb eines Jahres.

Institution →	A	B	C	D	E
OHNE SPRACHMODELL	36,7	26,1	43,3		
Philips Bigramm	61,5	47,4	65,6	68,3	
Philips Trigramm	63,0				
eigenes Bi-/Trigramm	61,8		67,1	71,4	58,4

Tabelle 1: Wortakkuratheitsrate zum SHORT-Korpus (in Prozent)

Institution →	A	B	C	D	E
OHNE SPRACHMODELL	39,3	27,1			
Philips Bigramm	61,7	47,3	60,2	66,6	
Philips Trigramm	62,5				
eigenes Bi-/Trigramm	62,2	26,5	60,9	68,6	54,3

Tabelle 2: Wortakkuratheitsrate zum LONG-Korpus (in Prozent)

5 Ausblick

Für den Sommer 1996 ist eine weitere Evaluierung der signalnahen Spracherkennung im Projekt Verbmobil vorgesehen. Die Evaluierung im Jahre 1994 und 1995 haben aber bereits gezeigt, daß die zentrale Auswertung der Erkennungsergebnisse, die Bereitsstellung der Evaluierungssoftware und die Festlegung von Trainings- und Testmaterial die Vergleichbarkeit von Erkennern für das Verbmobil-Szenario ermöglicht und die Einordnung der Erkennungsleistung der verschiedenen Systeme im nationalen Vergleich erlaubt. Ein Vergleich der Erkennungsergebnisse mit dem Jahr 1994 hat die erhebliche Steigerung in der Erkennungsleistung gezeigt.

Literatur

[OKUDA1976] T.Okuda, E.Tanaka & T.Kasai:*A Method for the Correction of Garbled Words Based on the Levenshtein 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

[PAULUS1994] Erwin Paulus, Michael Lehning: *Die Evaluierung von Spracherkennungssystemen in Deutschland*, Fortschritte der Akustik, DAGA 94- Dresden, Teil A, 147 - 156,