Regeladaptive kategoriale Analyse von Spontansprache - eine sprachenübergreifende Untersuchung

Nicole Beringer

Institut für Phonetik und Sprachliche Kommunikation, LMU München; Email: beringer@phonetik.uni-muenchen.de

1. Zusammenfassung

Die Untersuchung behandelt eine neue methologische Variante des MAUS-Systems zur Segmentierung[1] und Etikettierung[2] (S&E) von Spontansprache, mit der diese im Hinblick auf Vielfalt und Menge der vorhandenen Sprachkorpora qualitativ verbessert werden kann.

2. Nutzen der automatischen Segmentierung und Etikettierung in der Phonetik

Eine phonetische Analyse großer Korpora kann durch die Bereitstellung manueller S&E nicht erreicht werden[3].
Die Bereitstellung großer Mengen segmentierter und etikettierter Korpora, eine systematische Untersuchung phonetischer Realisierungen im Hinblick auf phonemische, akustische, perzeptive Variationen bis hin zum Sprachwandel, die Bildung von phonetisch manifestierbaren Ausspracheregeln sowie ein Einsatz der Segmente in der Sprachsynthese kann durch eine automatische S&E erfolgen. usw.

3. Das Münchner Automatische Segmentationssystem MAUS

Basierend auf den beiden Wissensquellen akustisch-phonetische Modelle (Segmentierung) und phonetisch-phonologische Ausspracheregeln (Etikettierung) wird bei der MAUS-Methode für die Generierung einer automatischen Transkription[4] auf breiter phonetischer Ebene[5] die S&E des Sprachsignals durch eine Viterbi-Suche im Variantengitter aller möglichen Aussprachen bestimmt.
Das Variantengitter wird durch die Anwendung von Ausspracheregeln auf die kanonische Form gebildet. Die Ausspracheregeln sind im klassischen Fall Expertenregeln, also eine "phonetisch-phonologische" Sammlung von einerseits phonologisch kategorisierten Aussprachevarianten, die von Experten nicht nur aus phonologischem Gesichtspunkt zusammengestellt, sondern andererseits zum Teil auch an konkretem Datenmaterial erstellt bzw. manifestiert wurde, also phonetisch basiert ist.
Alle Expertenregeln sind für die automatische Etikettierung gleich wahrscheinlich. Daher ist keine Aufteilung nach "Wichtigkeit" - v.a. auch im Hinblick auf die große Anzahl (ca. 5000 im Deutschen) der verschiedenen Regeln - möglich. Die Expertenregeln sind dadurch übergenerativ [3].
Eine bessere Übereinstimmung der automatischen Etikettierung durch die MAUS-Methode mit manuell erstellten Etikettierungen erreicht man im konkreten Fall durch die Einsetzung von statistischen Ausspracheregeln. Die statistischen Ausspracheregeln werden automatisch durch den Vergleich von Handetikettierungen mit der kanonischen Form, der Zitierform für die einzelnen Wörter oder Wortfolgen, gelernt. Gemäß der Auftretenshäufigkeit im handetikettierten Material werden die einzelnen Regeln statistisch gewichtet [2].

4. MAUS und andere Korpora

Eine automatische S&E von deutscher Spontansprache durch MAUS ist mit statistischen Regeln qualitativ hochwertig. Die Portierung auf andere Sprachen/Sprachstile ist jedoch sehr aufwendig, da die Verwendung von manuell etikettiertem Material voraussetzt, daß dieses existiert. Da die manuelle S&E von Spontansprache durch Experten jedoch sehr zeit- und kostenintensiv ist, ist manuell etikettiertes Material rar. Da zudem für die Erstellung statistischer Ausspracheregeln ca. eine Stunde handetikettiertes Trainingsmaterial vorhanden sein muß, um eine aussagekräftige Wahrscheinlichkeitsverteilung sowie eine gute Abdeckung der Aussprachevarianten zu gewährleisten, kann in kurzer Zeit kein ausreichendes Trainingsmaterial zur Verfügung gestellt werden.
Außerdem ist handsegmentiertes und -etikettiertes Material ebenfalls nicht "breit gestreut", was heißen soll, daß nicht für jeden Sprachstil manuell bearbeitetes Material vorliegt und somit das vorhandene "stilspezifisch" ist. Ein automatisches Etikettieren z.B. von dialektal gefärbter Spontansprache mit statistischen Ausspracheregeln trainiert auf schwach gefärbtem Material, wie es z.B. im Verbmobilkorpus [4] vorkommt, führt somit auch zu einer schlechten Übereinstimmung mit vorhandenen Handsegmentierungen.
Die Portierung des MAUS-Systems auf andere Sprachen, wie im Rahmen von Verbmobil auf Japanisch und Amerikanisch, oder andere Sprachstile, wie die dialektal gefärbten Daten des RVG-Korpus (Regional Variants of German) ist also in jedem Fall problematisch, da zwar Expertenregeln für Aussprachevarianten relativ einfach zu finden sind, diese aber statistisch nicht gewichtet sind und daher keine qualitativ zu Handsegmentierungen vergleichbare Etikettierung liefern. Zudem ist bereits vorhandenes manuell segmentiertes und etikettiertes Material für die Generierung der statistischen Ausspracheregeln zu "stilspezifisch" und daher nicht für andere Sprachstile und schon gar nicht für andere Sprachen verwendbar. Letztendlich ist stilbezogenes bzw. sprachenbezogenes handsegmentiertes Material meist nicht verfügbar.

5. MAUS-ER: Das MAUS-System erweitert auf andere Sprachen/Sprachstile

Zur S&E dialektal gefärbter Spontansprache, sowie von spontansprachlichen Äußerungen anderer Sprachen wurde daher ein Algorithmus entworfen und implementiert (MAUS-ER - MAUS-System ERweitert), der die oben genannten Probleme umgeht, die Effizienz der S&E durch statistische Gewichtung der Regeln jedoch weitgehend bewahrt.
Unter Verwendung einer ausreichenden Menge von Ausspracheregeln für Aussprachevarianten anstelle von Handsegmentierungen/-etikettierungen kann durch die iterative Einbindung dieser zunächst ungewichteten Regeln in einem Spracherkennungsprozeß eine äquivalente S&E entstehen, bei der die Effizienz durch statistische Gewich-

[1] Die Segmentierung von Sprache ist die Unterteilung des Sprachsignals in einzelne zeitlich sich nicht überlappende diskrete Abschnitte

[2] Die Etikettierung von Sprache ist demnach die Zuteilung eines Zeitsignals auf das Symbol (= der Inhalt einer bestimmten Klasse)

[3] Echtzeitfaktor manueller S&E von 400 bis 800 = pro Minute Spontansprache beträgt der Zeitaufwand für die Segmentierung ca. sechs bis zehn Stunden

[4] Die bei der Etikettierung entstandene Symbolfolge heißt Transkription oder allgemeiner die Verschriftung einer sprachlichen Äußerung.

[5] Die breite phonetische Ebene zeichnet sich durch folgendes Merkmal aus: die Symbole geben so weit wie möglich die phonetische Qualität einer Äußerung wieder, wobei die Symbole zu einem phonemischen Inventar gehören. Bei der **engen phonetischen Ebene** hingegen muß das Symbolinventar so gewählt werden, daß damit alle Sprachen der Welt in ihrer phonetischen Qualität transkribiert werden können.
Die Symbole der phonemischen Ebene bezeichnen die kleinsten segmentalen Lauteinheiten, die sog. Phoneme, die bedeutungsunterscheidend sind. Dieses Inventar ist im Großen und Ganzen sprachspezifisch.

tung der im Datenmaterial gesehenen Regeln jedoch weitgehend bewahrt wird.
Der geringe Anteil an erforderlichem handsegmentierten und -etikettierten Material zur Evaluierung ist relativ einfach und schnell zu erstellen. Basierend auf ungewichteten Expertenregeln oder plausiblen Regeln[6] werden iterativ statistisch gewichtete Regeln nach folgendem Prinzip erstellt.

- Gegeben sei eine ausreichende Menge von Expertenregeln für Aussprachevarianten des jeweiligen Sprachstils einer Sprache.
- Diese Expertenregeln werden iterativ in einem Spracherkennungsprozeß eingebunden, d.h. die zunächst ungewichteten Expertenregeln bekommen in dieser Phase eine statistische Gewichtung sofern sie im Datenmaterial gesichtet werden, ungesehene Regeln bleiben ungewichtet.
- Plausible neue Regeln aus der freien Erkennung, also Varianten, die gültige Silbenstrukturen der zu segmentierenden Sprache enthalten, werden ebenfalls statistisch gewichtet im Regelset integriert.
- Dieser Vorgang wird nun iterativ solange wiederholt, bis die automatische S&E unter Verwendung dieser iterativ gebildeten Regeln der Teststichprobe - dem manuell segmentierten und etikettierten Evaluationsmaterial, die maximale Übereinstimmung mit dem Evaluationsmaterial erhält (Konvergenz).

Die Anzahl der Regeln beeinflußt die Effizienz, d.h. bei zu wenig Regeln, daß die kanonische Form u. U. nicht modifiziert wird, bzw. statistisch ähnliche oder per Zufall auftretende Regeln beim iterativen Lernen zu vielen ähnlichen Regeln bzw. zu einer großen Verwechslungsmöglichkeit führen. Daher wird ein statischer oder dynamischer Schwellwert zur Begrenzung der Gesamtzahl der Regeln verwendet. Das Ausdünnen der Regeln erfolgt **nach** der Regelgenerierung der einzelnen Verarbeitungsschritte unter Verwendung einer a-priori Wahrscheinlichkeit.

6. Experimentelle Evaluation des Systems

Zum Training der akustischen Modelle diente das Hidden-Markov-Toolkit. Für Deutsch wurde handsegmentiertes Material (474 Turns des Kielkorpus) zum Training der HMMs verwendet, für Englisch mit 14875 Turns aus Verbmobil I und II[7] wurde ein "flat start" eingesetzt.
Das Training der statistischen Ausspracheregeln für den S&E Vergleich mit MAUS erfolgte durch den automatischen Vergleich der Handsegmentierungen des Kiel-Korpus (Deutsch), von Teilen der Verbmobil II-Teilkorpora Englisch und von Teilen der Verbmobil II-Teilkorpora Japanisch mit der kanonischen Form. Bei den MAUS-ER-Trainingssequenzen besteht dieselbe Aufteilung für Englisch und Japanisch. Die deutschen MAUS-ER-Trainingssequenzen setzen sich aus den Turns des Dialogs m112d der CD VM2.1 zusammen.
Die deutschen Verbmobildialoge der Teststichprobe bestehen aus 29953 Segmenten (in 328 Turns[7]) und die englischen Verbmobildialoge der Teststichprobe bestehen aus 5072 englischen Segmenten (in 137 Turns[7]) Die japanischen Verbmobildialoge der Teststichprobe bestehen aus 10766 japanische Segmente (in 239 Turns[7]).

Die nachfolgende Tabelle gibt einen detaillierten Einblick der Etikettierungsakkuratheiten zur manuellen Referenztranskription bei den verschiedenen Iterationsschritten ({mauser/free/prun}{init/1 [8]} sowie der kanonischen bzw. der MAUS-Transkription.

	Regelset	Deutsch	Englisch	Japanisch
kan-ref	kein	74,71%	34,24%	24,99%
MAUS-ref	statistisch	80,41%	40,69%	76,54%
MAUS-kan	statistisch	69,00%	40,69%	44,04%
mauserinit	Experten	74,08%	49,83%	63,69%
mauser1	iterativ	78,5%	49,83%	66,69%
freeinit	plausibel	57,64%	66,64%	66,64%
free1	iterativ plausibel	66,17%	68,32%	66,64%
pruninit	Experten	57,64%	49,83%	63,69%
	plausibel	52,12%	60,81%	62,54%
prun1	Experten	65,77%	66,64%	66,64%
	plausibel	58,82%	66,64%	66,64%

Tabelle 1. Evaluation des MAUS-ER-Algorithmus für die VM Korpora Deutsch, Englisch und Japanisch

Insgesamt zeigt sich, daß mit der iterativen Regeladaption die automatische Segmentierung und Etikettierung von Spontansprache im Hinblick auf die Vielfalt und der Menge der vorhandenen Sprachkorpora qualitativ verbessert werden kann.

7. Ausblicke

Eine egelbasierte Modellierung von Aussprachevarianten anstelle reiner akustisch-phonetischer Modelle wie sie hier vorgestellt wurde, kann neben ASR-Systemen auch in der Sprachsynthese eingesetht werden. Denn je besser die Segmente mit Handsegmentierungen übereinstimmen, desto weniger anfällig sind sie auf automatisch erzeugte Segmentierungen bezüglich inkonsistener Lautübergänge bzw. unnatürlich klingender Koartikulationseffekte.
Auch eine Verwendung von MAUS-ER-segmentierten Korpora bei phonetischen bzw. linguistischen Untersuchungenist vielversprechend. Automatisch segmentiertes und etikettiertes Material wurde z.B. bereits für die Erstellung einer Auftretensstatistik der deutschen Phoneme in gesprochener Sprache verwendet. Auch eine Analyse von linguistischen Phänomenen, wie z.B. mögliche Indizien für Sprachwandel, die Modellierung der regionalen Aussprachevariation oder die Weiterverarbeitung von Daten in der phonetischen Analyse zur konkatenativen Sprachsynthese bzw. zur Verbesserung von Sprachlernsoftware durch automatisch erstelltes S& E Materialist durchaus sinnvoll.

8. Bibliographie

[1] Nicole Beringer. Regeladaptive kategoriale Analyse von Spontansprache, Dissertation Juli 2002.

[2] A. Kipp. Automatische Segmentierung und Etikettierung von Spontansprache. *Shaker Verlag*, Aachen, 1999.

[3] M.-B. Wesenick. *Entwurf eines unterspezifizierenden Regelsystems der Aussprache des Deutschen als Basis f"ur empirische Untersuchungen*. Magisterarbeit, Ludwig-Maximilians-Universit"at M"unchen, 1994.

[4] K. Weilhammer, F. Schiel, U. Reichel. Multi-Tier Annotations in the Verbmobil Corpus. In *Proc. of the LREC 2002*, to appear.

[6] *Plausible Regeln sind Regeln, die aus der freien Phonemerkennung gebildet wurden und Veränderungen beschreiben, die entweder in Expertenregeln bzw. in statistischen Regeln der zu etikettierenden Sprache vorhanden sind.* Dabei wird nicht das Phonem an sich betrachtet, sondern dessen phonologische Klasse, wie zum Beispiel:
Die freie Erkennung liefert folgende Regel fürs Deutsche: a-k-t> a-x-t
Es gibt die statistische oder Experten-Regel fürs Deutsche: E-k-t> E-x-t
Die Regel aus der freien Erkennung wäre plausibel, da der Linkskontext der Regel zur selben Klasse von Phonemen wie der Linkskontext der statistischen bzw. Experten-Regel gehört, nämlich zu den stimmhaften Plosiven. Der Rechtskontext ist in diesem Beispiel sogar identisch, ebenso wie die sich verändernde Phonemfolge.

[7] genaue Aufteilung siehe [1]

[8] 1. Iteration