AUTOMATISCHE ZEICHENSETZUNG IN SPRACHERKENNUNGSSYSTEMEN: ENTSCHEIDUNGSBAUM UND SPRACHMODELL IM VERGLEICH

Heike Adel, Kevin Kilgour, Sebastian Stüker, Alex Waibel

InterACT, Institute of Anthropomatics
Karlsruher Institut für Technologie (KIT)

Kurzfassung: In diesem Artikel wird die Möglichkeit vorgestellt, Spracherkennerausgaben in englischer Sprache durch Punkte und Kommata automatisch zu strukturieren.
Dazu werden ein Entscheidungsbaum auf prosodischen Merkmalen und Wortarten und ein Hidden-N-Gramm-Modell auf Worten trainiert. Für die Entscheidung, ob und welches Satzzeichen gesetzt werden sollte, werden die Wahrscheinlichkeiten der Modelle linear interpoliert.
Als prosodische Merkmale werden Pausen nach dem aktuellen Wort, eine Wortlängen-Wortsprechdauer-Relation sowie die Information, ob nach dem aktuellen Wort ein Sprecherwechsel vorliegt, betrachtet. Die Arbeit zeigt, dass die prosodischen Merkmale des Entscheidungsbaums allein ähnliche Ergebnisse liefern wie die Worte des Sprachmodells. Allerdings wird auch deutlich, dass eine Modellkombination zu besseren Ergebnissen führt.
Die verwendeten Modelle setzen die Zeichen, die sie finden, zufriedenstellend, erkennen aber zu wenige Zeichen. Daher wird ein Faktor eingeführt, der dafür sorgt, dass die Wahrscheinlichkeit für „kein Zeichen“ heruntergewichtet wird und die Wahrscheinlichkeiten für die Satzzeichen entsprechend hochgewichtet werden. Dieser Faktor berechnet sich aus einer linearen Gleichung, die von der Anzahl der Worte seit dem letzten Satzzeichen abhängt.
Der Ansatz führt zu einer Fehlerrate bei der Satzgrenzenerkennung von 65,95 % auf den Hypothesen des Spracherkenners sowie zu einer Fehlerrate von 45,83 % auf den Referenztexten.

1 Einführung

1.1 Motivation

Automatische Spracherkennung durchdringt unseren Alltag zunehmend. Das Lesen der Ausgabe eines Spracherkenners ist aber aufgrund der fehlenden Satzzeichen oft schwer.
Das folgende Beispiel aus einem Podcast der BBC vom 10.03.2010 verdeutlicht dies:

> you see Bob Geldof in that clip that you played earlier on said there was not a single shred of evidence of a diversion of funds
>
> You see, Bob Geldof, in that clip that you played earlier on, said there was not a single shred of evidence of a diversion of funds.

Ein angenehmes Lesen von Spracherkennerausgaben kann die Akzeptanz von Spracherkennern deutlich steigern. Auch im Zuge der Inbetriebnahme des Lecture Translators am KIT ist dies ein wichtiger Aspekt. Der Lecture Translator soll vor allem ausländischen Studierenden helfen,

den Vorlesungen zu folgen, indem diese in Echtzeit transkribiert und übersetzt werden [4, 10]. Eine Strukturierung des Textflusses erleichtert auch maschinelle Übersetzung und automatische Textzusammenfassung.

1.2 Verwandte Arbeiten

Bisherige Arbeiten auf dem Gebiet der automatischen Zeichensetzung lassen sich in zwei Gruppen teilen: Zum Einen werden nur textuelle Merkmale betrachtet [6, 12, 13], zum Anderen auch prosodische Merkmale hinzugezogen, um Satzgrenzen zu erkennen [2, 3, 5, 7, 8, 17].
Die Zahl der verwendeten Modelle reicht von N-Gramm-Modellen [2, 6, 8, 17] über Conditional Random Field Modelle [11, 12, 13], Maximum Entropie-Modelle [7, 11] bis hin zu Hidden Markov Modellen [2, 8, 11, 17] und Entscheidungsbäumen [2, 8, 17].
In dieser Arbeit wird ebenso wie in der Arbeit von [8] ein Entscheidungsbaum auf prosodischen Merkmalen sowie Wortarten trainiert und mit einem Hidden-N-Gramm-Sprachmodell interpoliert. Die Interpolation wird hier jedoch noch um eine Wahrscheinlichkeitsanpassung erweitert. [5] verwendet ebenfalls die Modelle Entscheidungsbaum und N-Gramm, nutzt aber mathematisch und heuristisch motivierte Formeln zur Kombination.

1.3 Besondere Schwierigkeiten

Besondere Schwierigkeiten ergeben sich aus der Umgebung des Lecture Translators. Diese umfasst frei gesprochene Sprache (in eventuell geräuschvollen Umgebungen) und Spracherkennung und -übersetzung in Echtzeit.

1.3.1 Frei gesprochene Sprache

Frei gesprochene Sprache erschwert die Spracherkennung, da sie Unregelmäßigkeiten, Denkpausen und Selbstverbesserungen aufweist [5, 7, 8, 11]. Des Weiteren sind Phrasengrenzen in geschriebenen Texten nicht denen in frei gesprochener Sprache gleichzusetzen: Im Allgemeinen sind gesprochene Phrasen (Sentence-like Units, SUs) kürzer [8, 11]. Dieses Wissen ist für die korrekte Interpretation von Fehlerraten nötig. Zudem gibt es in frei gesprochenen Texten auch unvollständige SUs, wenn der Sprecher sich selbst unterbricht oder unterbrochen wird.

1.3.2 Online-Erkennung, keine Vorsegmentierung

Die meisten Ansätze, Satzzeichen zu erkennen, befassen sich mit dem Fall, dass der ganze Text bekannt ist und so als Kontext auch Worte berücksichtigt werden können, die nach der zu untersuchenden Stelle stehen. Online-Erkennung erschwert das korrekte Einfügen von Satzzeichen aufgrund des geringeren Kontextes für ein Wort. [2] und [16] simulieren Online-Zeichensetzung, indem sie den Kontext nur auf die Vergangenheit beschränken. Im Unterschied dazu kann im Lecture-Translator als Kontext der Block verwendet werden, den der Spracherkenner auf einmal verarbeitet und ausgibt. Gerade zu Beginn des Blocks ist demnach auch rechtsseitiger Kontext vorhanden.
Aufgrund der Online-Erkennung sind weder Vorsegmentierungen noch mehrere Dekodierungsdurchläufe möglich. Segmentierungsfehler und damit falsch erkannte Pausen oder Worte führen zu falschen Eingaben für die Modelle und damit zu Zeichensetzungsfehlern.

2 Versuchsaufbau

Die Zeichensetzung wird als Klassifizierungsproblem betrachtet: Je nachdem, ob einem Wort ein Komma, Satzendezeichen oder kein Zeichen folgt, wird es in eine der Klassen „COMMA“, „END“ oder „NONE“ eingeordnet.
Die verwendeten Modelle wurden nicht zur Unterscheidung zwischen Punkt und Fragezeichen trainiert, weil der verwendete Kontext meist zu gering ist, um einen langen Fragesatz als solchen zu erkennen [6].
Diese Arbeit verwendet sowohl textuelle als auch prosodische Merkmale. Unter Prosodie versteht man Betonung, Sprechrhythmus und Sprechmelodie. Diese Merkmale finden sich vor allem in der Länge von Pausen und Wörtern, im Verlauf der Sprechfrequenz und in der Lautstärke der Sprache wieder [15, 19].
Für die textuellen Merkmale wird ein N-Gramm-Modell trainiert, für die prosodischen Merkmale ein Entscheidungsbaum. Die Klassifikationen dieser Modelle werden im Anschluss anhand ihrer Wahrscheinlichkeiten interpoliert.

2.1 Hidden-N-Gramm-Modell

Es wird ein Hidden-N-Gramm-Modell mit Kontextlänge 4 aufgebaut. Diese kann noch relativ robust trainiert werden und führt durch mehr Hintergrundwissen zu reduzierten Fehlerraten.
Bei Hidden N-Grammen gibt es zusätzlich zum Vokabular des Spracherkenners ein weiteres aus versteckten Elementen (den Satzzeichen). Beim Training werden diese ebenfalls beachtet. Ein Text ohne Satzzeichen kann dann mit diesen versehen werden, indem das N-Gramm-Modell als Hidden Markov Modell genutzt wird [19]. Zum Bau des Hidden-N-Gramm-Modells wird das SRILM-Toolkit verwendet [18].

2.2 Entscheidungsbaum

Der Entscheidungsbaum wird auf folgenden Merkmalen trainiert:

- Pause nach dem aktuellen Wort
- Sprecherwechsel nach aktuellem Wort
- Wortdauer-Wortlänge-Relation des aktuellen Wortes (zeigt an, ob ein Sprecher besonders langsam gesprochen hat und dient damit der Erkennung von Phrasenenden)
- vorherige und nachfolgende Wörter mit einer bestimmten Kontextlänge
- Wortarten (Part-of-speech-Tags, POS) der vorherigen und nachfolgenden Wörter mit gleicher Kontextlänge
- vor dem aktuellen Wort gesetztes Satzzeichen

Alle Wörter des Vokabulars in den Entscheidungsbaum aufzunehmen, erweist sich sowohl bezüglich des Speicherplatzes als auch bezüglich der Rechenzeit als ineffizient und nicht für Echtzeiterkennung geeignet.
Als Parameter ergeben sich damit die Länge des links- und rechtsseitigen Kontexts und die Anzahl aufgenommener Wörter. Zur Schätzung der Parameter wurden diverse Versuche durchgeführt, die in Abschnitt 2.5 zusammengefasst dargestellt werden.
Zum Bau des Entscheidungsbaums wird die Implementierung von C4.5 der University of Regina verwendet [1]. Dieser Algorithmus ist für diese Arbeit besser geeignet als der Algorithmus ID3, da er die Möglichkeit bietet, Werte unbesetzt zu lassen. In der Online-Satzzeichenerkennung ist es so auch möglich, Wahrscheinlichkeiten für Fälle zu berechnen, in denen nicht der gesamte Kontext bekannt ist.

[1] http://www2.cs.uregina.ca/ dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html

Für die Bestimmung der Wortarten wird der Part-of-speech-Tagger der Universität Stuttgart verwendet [2]. Im Englischen basiert dieser auf dem Penn-Treebank-Tagset [3].

2.3 Kombination

Für die Klassifikation werden die Wahrscheinlichkeiten der beiden vorgestellten Modelle nach Gleichung 1 linear interpoliert.

$$P(Klasse) = \lambda \cdot P_{Entscheidungsbaum}(Klasse) + (1-\lambda) \cdot P_{N-Gramm}(Klasse) \tag{1}$$

An den niedrigen Recallwerten der beiden Modelle wird deutlich, dass zu wenig Zeichen gesetzt werden (vergleiche Abschnitt 3). Daher wird pro Modell die Wahrscheinlichkeit für die Klasse „NONE“ wie in Gleichung 2 beschrieben heruntergewichtet und die freiwerdende Wahrscheinlichkeitsmasse auf die beiden anderen Klassen „COMMA“ und „END“ verteilt.

$$P(NONE)_{neu} = P(NONE)_{alt} \cdot \alpha(k) \tag{2}$$

Der Herabgewichtungsfaktor α berechnet sich durch eine lineare Gleichung in Abhängigkeit der Anzahl Wörter k seit dem letzten Satzzeichen:

$$\alpha(k) = \begin{cases} 0 & a \cdot k + b > 1 \\ 1 - (a \cdot k + b) & sonst \end{cases} \tag{3}$$

Hintergrund dieser Formel ist die Beobachtung, dass Sätze begrenzte Längen besitzen. Je mehr Worte bereits in einem Satz vorhanden sind, desto mehr steigt daher die Wahrscheinlichkeit, bald ein Satzzeichen zu sehen.
Die Kombination der Modelle ergibt fünf weitere Parameter, die im Folgenden untersucht werden: a und b für die Umgewichtung von N-Gramm-Modell und Entscheidungsbaum sowie den Interpolationsfaktor λ.

2.4 Trainingsdaten und Testdaten

2.4.1 Trainingsdaten

Zum Trainieren des Sprachmodells werden Broadcast-News und Zeitungstexte verwendet. Die verwendeten Zeitungstexte stammen aus der Central News Agency of Taiwan, der Los Angeles Times, der Washington Post und der New York Times. Sie sind Teil der Gigaword-Daten [14]. Die Texte umfassen ungefähr 3,6 Milliarden Wörter.
Der Entscheidungsbaum wird mit den Quaero-Trainingsdaten für Zeichensetzung von 2011 [9] sowie den allgemeinen akustischen Trainingsdaten von Quaero aus dem Jahr 2010 trainiert [20]. Diese Daten umfassen ungefähr 1,2 Millionen Wörter.

2.4.2 Testdaten

Die Evaluierung erfolgt auf den Quaero-Development-Daten von 2010. Bezüglich der Zeichensetzung liegen zwei Versionen vor. Diese wurden von verschiedenen Muttersprachlern mit linguistischem Hintergrund erstellt [9].
Tabelle 1 zeigt die Verteilung der Satzzeichen sowie die Gesamtzahl der Worte des Textes.
Version 1 enthält mehr Satzzeichen als Version 2. Ein Inter-Annotator Agreement (IAA) zeigt,

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.pdf

dass 4225 Zeichen übereinstimmen. Des Weiteren setzt die zweite Version 644 Zeichen, die in Version 1 nicht vorkommen und Version 1 enthält 704 Zeichen, die Version 2 nicht beinhaltet. An 452 Stellen steht zwar in beiden Versionen ein Satzzeichen, allerdings nicht das gleiche. Bei der Bewertung der Modelle werden beide Versionen betrachtet: Ein Zeichen wird als korrekt gewertet, wenn es von mindestens einer der Referenzen bestätigt wird.

Tabelle 1 - Versionen der Development-Daten

Version	Worte	Satzenden (Punkte und Fragezeichen)	Kommata
Version 1	39167	2331(5,24 %)	3050 (6,85 %)
Version 2	39167	2308 (5,19 %)	3013 (6,77 %)

2.5 Parameterschätzungen

2.5.1 Bewertungsmaßstäbe

Zur Evaluierung werden Einfügungsfehler I, Auslassungsfehler D und Verwechslungsfehler S sowie die Anzahl der korrekten Satzzeichen C berechnet. Da der Fokus auf der generellen Strukturierung des Textes durch Zeichen und nicht in erster Linie auf der Korrektheit der Zeichen selbst liegt, werden SU-Precision, SU-Recall und SU-Fehlerrate betrachtet [5, 11, 13].

$$P_{SU} = \frac{C+S}{C+S+I} \quad (4) \qquad R_{SU} = \frac{C+S}{C+S+D} \quad (5) \qquad SU-Error = \frac{I+D}{C+S+D} \quad (6)$$

Entscheidungen für oder wider einen Parameter werden auf Basis der SU-Fehlerrate getroffen.

2.6 Parameter des Entscheidungsbaums

2.6.1 Kontextlänge

Der Entscheidungsbaum wird zunächst mit einer festen Schlüsselwortzahl auf verschiedene Kontexte (bis zu 6 Wörter in die Vergangenheit und Zukunft) getestet. Dabei ergeben sich die besten SU-Fehlerraten bei einem Kontext von 3 in die Vergangenheit und 1 in die Zukunft. Bei großen Kontexten sind die Fehlerraten deutlich höher als bei kleineren.

2.6.2 Schlüsselwortanzahl

Im nächsten Schritt wird der Baum mit dem besten Kontext mit verschiedenen Schlüsselwörtern getestet. Es werden jeweils die x häufigsten Wörter, die in den Trainingstexten vor oder nach einem Satzzeichen vorkamen, gewählt ($x \in \{0, 125, 250, 500, 1000\}$). Alle anderen Worte werden auf einen Platzhalter abgebildet. Es wird zwischen absoluter und relativer Häufigkeit unterschieden. Bei relativer Häufigkeit wird die Anzahl der Vorkommnisse vor oder nach einem Satzzeichen noch durch die Gesamtanzahl der Vorkommnissen im gesamten Text dividiert. Auf diese Art erhält eine Konjunktion wie „because“ eine höhere Wertung als ein Wort wie „you“, da „because“ im Gegensatz zu „you“ fast nur im Zusammenhang mit Satzgrenzen zu finden ist. Die absolute Zählweise führt dennoch zu niedrigeren Fehlerraten als die relative. Die geringste Fehlerrate liegt bei 125 Worten mit absoluter Zählweise vor. Der Grund für die besseren Werte bei Aufnahme eher weniger Worte könnte darin liegen, dass viele Wörter im Training nicht oft genug gesehen werden, um für sie aussagekräftige Wahrscheinlichkeiten schätzen zu können.

2.6.3 Resultierender Entscheidungsbaum

Abbildung 1 zeigt den aus den Versuchen resultierenden Entscheidungsbaum. Es ist erkennbar, dass die ersten und damit wichtigsten Fragen die Prosodie betreffen.

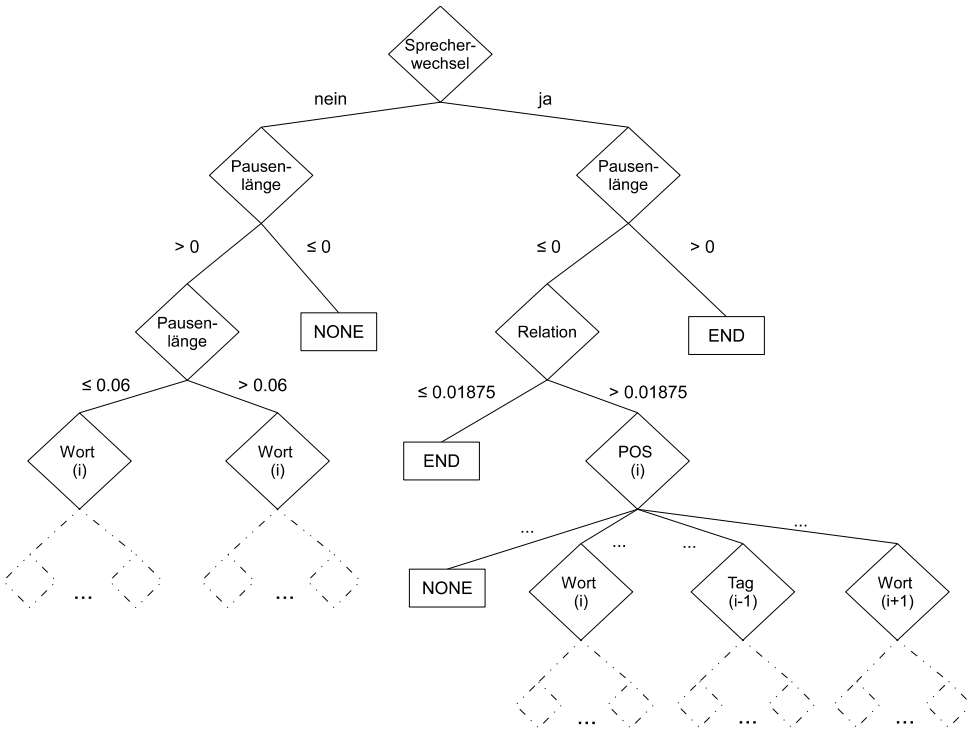

Abbildung 1 - Resultierender Entscheidungsbaum

2.7 Parameter der Kombination der Modelle

Wie in Abschnitt 2.3 beschrieben wird der Entscheidungsbaum mit dem N-Gramm-Modell kombiniert. Für die Umgewichtung werden für beide Modelle in 0,2er Schritten alle möglichen Werte zwischen 0 und 1 getestet. Im Anschluss daran werden für die besten Umgewichtungsfaktoren in 0,1er Schritten alle möglichen Interpolationsfaktoren zwischen 0 und 1 probiert.
Während der Entscheidungsbaum bei einer Umgewichtung, die von der Anzahl bisher gesehener Worte seit dem letzten Satzzeichen abhängt, bessere Ergebnisse liefert, zeigt sich beim N-Gramm-Modell, dass dieses mit einem konstanten Faktor ($a = 0$) umgewichtet werden sollte. Die konkreten Ergebnisse sind $a = 0{,}6$ und $b = 0{,}4$ für den Entscheidungsbaum sowie $a = 0$ und $b = 0{,}2$ für das Sprachmodell.
Als Interpolationsfaktor λ ergibt sich ein Wert von 0,3.
Dass der Entscheidungsbaum dadurch mit weniger als 50% einfließt, erklärt sich aus der Tatsache, dass er deutlich stärker umgewichtet wird als das N-Gramm-Modell. Würde er ein höheres Gewicht bekommen, entstünden viele Einfügungsfehler.

3 Ergebnisse

In diesem Abschnitt werden die konkreten Ergebnisse des Zeichensetzungsmodells vorgestellt. Als Baseline dient die Komponente des Lecture Translators, die den erkannten Textstrom auf Basis von Sprachmodellinformationen in Sätze zerlegt.

3.1 Ergebnisse der einzelnen Modelle

An den in Tabelle 2 dargestellten Ergebnissen wird deutlich, dass eine Kombination von N-Gramm-Modell und Entscheidungsbaum die Satzgrenzenerkennungsrate deutlich erhöht.
Das Baseline-Modell erweist sich auf den gleichen Testdaten in allen Bewertungsmaßstäben weniger gut als das neu entwickelte Modell. Da das Baseline-Modell eher für die Erkennung von

Satzenden optimiert wurde, die verwendeten Bewertungsmaßstäbe aber auch gegen Kommata scoren, wird zum besseren Vergleich das neue Modell auch so ausgeführt, dass es nur Punkte setzt und Kommata überspringt. (In der Tabelle wird dies als „neues Modell ohne Kommata“ bezeichnet.) Nicht gesetzte Kommata werden dann auch nicht als Fehler gewertet.

Tabelle 2 - Modellergebnisse in Prozent

	SU-Precision	SU-Recall	SU-Fehlerrate
N-Gramm-Modell alleine	59,53	22,43	92,82
Entscheidungsbaum alleine	74,12	16,35	89,36
Kombination	69,96	59,66	65,95
Baseline Modell	62,15	72,23	71,77
neues Modell ohne Kommata	69,32	64,63	63,97

3.2 Test auf Referenztexten und Spracherkennertranskriptionen

Ebenso wie [11] und [19] zeigt auch Tabelle 3, dass die Satzgrenzenerkennung auf Referenztexten bessere Ergebnisse liefert als auf den Hypothesentexten des Spracherkenners. Das Zeichensetzungssystem leidet an den Fehlern des Spracherkenners.

Tabelle 3 - Fehlerraten des Systems auf dem Referenztext im Vergleich zu Spracherkennerausgaben (in Prozent)

	SU-Precision	SU-Recall	SU-Fehlerrate
Spracherkennerausgabe	69,96	59,66	65,95
Referenztext	79,41	73,14	45,83

3.3 Interpretation der Fehlerraten

Wie eingangs erwähnt ergeben sich besondere Schwierigkeiten aus frei gesprochener Sprache. Gesprochene Phrasen (SUs) können nicht eins-zu-eins auf geschriebene Sätze abgebildet werden. Des Weiteren wird in den Versuchen keinerlei Vorsegmentierung verwendet. Dass ein großer Teil der Satzgrenzenerkennungsfehler von Segmentierungsfehlern des Spracherkenners herrührt, zeigt der Vergleich mit den Referenztexten. Aber auch dieser Test ist nicht frei von Segmentierungsfehlern: Es werden zwar die korrekten Wörter verwendet, die Pauseninformationen stammen aber immer noch aus den Segmentierungen des Spracherkenners.

4 Fazit und Ausblick

Das vorgestellte System dient der automatischen Strukturierung von Spracherkennerausgaben. Aus der Verbesserung der Ergebnisse durch Kombination der verwendeten Modelle kann man schließen, dass sowohl textuelle als auch prosodische Merkmale von großer Bedeutung sind.
Ein Grund für die hohe Auslassungsrate von Satzzeichen könnte in mangelnden Trainingsdaten liegen. Bei der Auswahl des Trainingsmaterials sollte allerdings darauf geachtet werden, dass die Zeichensetzung konsistent ist. Quaero versucht, durch die Erstellung von Richtlinien eine konsistente Zeichensetzung zu erreichen. Allerdings enthalten die Trainingsdaten für das erstellte System auch Texte, die diesen Richtlinien nicht entsprechen. Als Beispiel seien die Zeitungsartikel genannt, mit denen das Sprachmodell trainiert wurde.
Eine weitere Möglichkeit, den Entscheidungsbaum zu verbessern, liegt in der Hinzunahme weiterer prosodischer Merkmale. Viele Arbeiten zeigen, dass der Verlauf der Sprechfrequenz neben der Pause ein sehr verlässliches Merkmal für das Erkennen von Satzgrenzen darstellt. Zudem

kann die Sprechfrequenz der Fragezeichenerkennung dienen. Denn der Frequenz- (und damit der Tonhöhen-) Verlauf ist bei einem Fragesatz charakteristisch anders als bei einem Aussagesatz.

Literatur

[1] ADEL, H.: *Automatische Zeichensetzung in Spracherkennungssystemen - Entscheidungsbaum und Sprachmodell im Vergleich*. Bachelorarbeit, KIT (Karlsruher Institut für Technologie), 2011.

[2] BARON, D., E. SHRIBERG und A. STOLCKE: *Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues*. In: *Seventh International Conference on Spoken Language Processing*, 2002.

[3] CHRISTENSEN, H., Y. GOTOH und S. RENALS: *Punctuation annotation using statistical prosody models*. In: *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, 2001.

[4] FÜGEN, C., M. KOLSS, M. PAULIK, S. STÜKER, T. SCHULTZ und A. WAIBEL: *Open Domain Speech Translation: From Seminars and Speeches to Lectures*. TC-STAR workshop on speech to speech translation, Barcelona, Spain, pp.81-86, 2006.

[5] GOTOH, Y. und S. RENALS: *Sentence boundary detection in broadcast speech transcripts*. 2000.

[6] GRAVANO, A., M. JANSCHE und M. BACCHIANI: *Restoring punctuation and capitalization in transcribed speech*. In: *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, S. 4741–4744. IEEE, 2009.

[7] HUANG, J. und G. ZWEIG: *Maximum entropy model for punctuation annotation from speech*. In: *Seventh International Conference on Spoken Language Processing*, 2002.

[8] KIM, J., S. SCHWARM und M. OSTENDORF: *Detecting structural metadata with decision trees and transformation-based learning*. In: *Proc. of HLT/NAACL*, S. 137–144, 2004.

[9] KOLÁŘ, J. und L. LAMEL: *On Development of Consistently Punctuated Speech Corpora*. In: *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[10] KOLSS, M., M. WÖLFEL, F. KRAFT, J. NIEHUES, M. PAULIK und A. WAIBEL: *Simultaneous german-english lecture translation*. In: *Proc. of International Workshop on Spoken Language Translation*, 2008.

[11] LIU, Y., E. SHRIBERG, A. STOLCKE, D. HILLARD, M. OSTENDORF und M. HARPER: *Enriching speech recognition with automatic detection of sentence boundaries and disfluencies*. Audio, Speech, and Language Processing, IEEE Transactions on, 14(5):1526–1540, 2006.

[12] LU, W. und H. NG: *Better punctuation prediction with dynamic conditional random fields*. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, S. 177–186. Association for Computational Linguistics, 2010.

[13] MORGAN, W.: *Sentence Unit Detection without an Audio Signal*. 2009.

[14] PARKER, R., D. GRAFF, J. KONG, K. CHEN und K. MAEDA: *English Gigaword Fourth Edition*. Philadelphia, 2009. Linguistic Data Consortium.

[15] SELKIRK, E.: *Sentence prosody: Intonation, stress, and phrasing*. 1995.

[16] SHRIBERG, E., A. STOLCKE und D. BARON: *Can Prosody Aid the Automatic Processing of Multi-Party Meetings? Evidence from Predicting Punctuation, Disfluencies, and Overlapping Speech*. In: *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, 2001.

[17] SHRIBERG, E., A. STOLCKE, D. HAKKANI-TÜR und G. TÜR: *Prosody-based automatic segmentation of speech into sentences and topics*. Speech communication, 32(1):127–154, 2000.

[18] STOLCKE, A.: *SRILM-an extensible language modeling toolkit*. In: *Seventh International Conference on Spoken Language Processing*, 2002.

[19] STOLCKE, A., E. SHRIBERG, R. BATES, M. OSTENDORF, D. HAKKANI, M. PLAUCHE, G. TUR und Y. LU: *Automatic detection of sentence boundaries and disfluencies based on recognized words*. In: *Fifth International Conference on Spoken Language Processing*, 1998.

[20] STÜKER, S., K. KILGOUR und F. KRAFT: *Quaero 2010 Speech-to-Text Evaluation Systems*. High Performance Computing in Science and Engineering'11, S. 607–618, 2012.