Eine Infrastruktur zur Erforschung multimodaler Kommunikation

Uhrig, Peter
peter.uhrig@uos.de
Universität Osnabrück, Deutschland

Eine Infrastruktur zur Erforschung multimodaler Kommunikation

Dieser Vortrag zeigt, wie mit Hilfe des Distributed Little Red Hen Lab eine umfassende Datenbank und Forschungsinfrastruktur geschaffen wurde (und immer noch wird), mit der sich viele Aspekte multimodaler Kommunikation auf Basis der Fernsehaufnahmen des NewsScape-Projekts untersuchen lassen (Steen/Turner 2013).

Datensammlung und Datenbasis

Mit dem UCLA Library Broadcast NewsScape steht Forschern mittlerweile eine Datensammlung von über 400.000 Stunden digitaler Fernsehaufnahmen aus über 10 Jahren zur Verfügung, nicht nur vom US-amerikanischen Fernsehmarkt, aber mit einem starken Fokus auf demselben (vgl. Tabellen 1 und 2). Aufgrund besonderer Einschränkungen des Urheberrechts in den USA darf ein Archiv oder eine Bibliothek „News“ aufnehmen und Forschern zur Verfügung stellen. Bei NewsScape wird der Begriff *News* relativ weit ausgelegt, so dass sich auch politische Comedy oder verschiedenste Talkshows in den Aufnahmen finden. Aufgrund gesetzlicher Vorgaben müssen in den USA alle Sendungen mit Untertiteln ausgestrahlt werden, die NewsScape ebenfalls mit aufnimmt, so dass sofort eine relativ brauchbare Verschriftlichung vorliegt, wodurch die Aufnahmen durchsuchbar werden. Dies betrifft auch in den USA ausgestrahlte spanischsprachige Sendungen. Insgesamt nimmt Red Hen Sendungen in mehr als 15 Sprachen auf, unter anderem inzwischen auch in China und Indien, wobei nicht in allen Sprachen bzw. nicht auf allen Sendern Untertitel ausgestrahlt werden. Das Distributed Little Red Hen Lab kooperiert mit dem Open-Source-Projekt CCExtractor um verschiedenste technische Umsetzungen von Untertiteln in Text umwandeln zu können.

Video	
Videodateien	451.974
Laufzeit in Stunden	350.223
Text	
Untertiteldateien	452.208
OCR-Dateien (für Text im Bild)	428.920
TPT-Dateien (heruntergeladende Transkripte)	37.148
Wörter in Untertiteldateien	2,82 Mrd.
Wörter in OCR-Dateien	981,54 Mio.
Wörter in TPT-Dateien	440,38 Mio.
Bilder	
Miniaturbilder	126,08 Mio.

Tabelle 1: Zahlen zur gesamten Sammlung, Mitte November 2017 (übersetzt aus Uhrig 2018)

Sprache	Laufzeit	Wörter	Kommentar
Amerikanisches Englisch	298,004:48:10	2,089,518,746	
Spanisch	15,104:47:23	78,075,367	ca. 60% mexikanisches Spanisch
Französisch	11,425:36:07	8,222,300	verschiedene Varitäten; viele Aufnahmen ohne Untertitel
Internationales Englisch	8,271:55:02	35,646,649	Al Jazeera, France 24, Deutsche Welle, Russia Today, …
Persisch	5,103:04:54	0	Übertragung ohne Untertitel
Norwegisch	3,241:49:55	7,466,801	Aufnahmen seit 2007, Untertitel seit 2012
Britisches Englisch	2,313:59:54	14,545,895	
Russisch	1,905:47:52	6,511,767	
Deutsch	1,362:15:13	6,381,895	
Schwedisch	1,017:41:15	1,661,240	Aufnahmen seit 2011, Untertitel seit 2015
Portugiesisch	873:31:57	4,897,107	
Dänisch	866:47:26	4,628,942	
Niederländisch/ Flämisch	565:56:41	4,363,813	
Tschechisch	413:47:34	2,956,235	
Polnisch	262:57:42	1,672,483	
Arabisch	148:51:14	0	Übertragung ohne Untertitel

Tabelle 2: Verteilung auf die einzelnen Sprachen (übersetzt aus Uhrig 2018)

Inzwischen zeigt sich jedoch, dass z.T. die Ergebnisse automatischer Spracherkennung im Englischen näher am gesprochenen Wort liegen als die Untertitel, so dass mittelfristig davon auszugehen ist, dass das Vorhandensein von Untertiteln an Relevanz verliert.

Datenverarbeitung

In einem seit 2014 laufenden Projekt wird die Datensammlung auch für die linguistische Forschung aufbereitet, so dass sie nicht nur für traditionell linguistische Fragestellungen sondern auch für multimodale Forschung genutzt werden kann (momentan vorrangig für das Englische).

Textbasierte maschinelle Sprachverarbeitung

Amerikanische Untertitel werden fast ausschließlich in Großbuchstaben ausgestrahlt. Dies stellt eine nicht zu unterschätzende Herausforderung für die weitere Verarbeitung dar, da viele Natural Language Processing (NLP)-Werkzeuge massiv schlechtere Ergebnisse liefern, wenn das Eingabeformat nur aus Großbuchstaben besteht. Das bereits fängt mit der Satzsegmentierung an, für die (gerade im Englischen) die Großschreibung am Satzanfang ein wichtiger Hinweis darauf ist, wo Satzgrenzen zu finden sind. Ein neu entwickelter Satztrenner speziell für die Untertiteldaten und ihre Besonderheiten – Zeilen sind 32 Zeichen lang; Sätze beginnen oft auf einer neuen Zeile – verbesserte die Ergebnisse deutlich, was auch für nachfolgende Verarbeitungsschritte, vor allem das syntaktische Parsing, vorteilhaft ist.

Auch beim Part-of-Speech Tagging zeigte sich, dass reine Großbuchstaben zu schlechten Ergebnissen führen. Das „caseless“ Modell von Stanford CoreNLP (Manning et al. 2014) für das Englische sorgte hier für gute Ergebnisse, die den Ergebnissen für Text mit Groß- und Kleinschreibung kaum nachstehen. Zusätzlich kann man optional ein Modul namens „TrueCase“ nachschalten, das versucht, auf Basis der PoS tags die ursprüngliche Groß- und Kleinschreibung zu erraten.

Für das syntaktische Parsing bietet Stanford ebenfalls ein „caseless“-Modell an, das jedoch auf relativ alter Parsertechnologie aufbaut (klassischer PCFG-Parser mit regelbasiertem Konverter für Dependenzen) und bei unseren Tests auf englischen Daten mit normaler Groß- und Kleinschreibung deutlich schlechter abschneidet das aktuelle Modell (Chen and Manning 2014) namens *dependency neural network* (F-Score labeled attachment: 76,22 vs. 79,56). Es war also notwendig, eine genaue Evaluation durchzuführen, um die bestmögliche Parameterkombination zu ermitteln. Insgesamt wurden 576 Parameterkombinationen evaluiert. Dazu wurde das Korpus ANC MASC mit verschiedenen Parsern und Modellen sowie mit der Originalschreibweise, nur Großschreibung, nur Kleinschreibung und mit den Ergebnissen des TrueCase-Moduls (wo das möglich war) geparst. Es zeigte sich, dass mit TrueCase das Ergebnis des modernen *dependency neural network* Parsers aus Stanford CoreNLP die Ergebnisse relativ nahe an den Ergebnissen mit Originalschreibweise lagen (79,18 vs. 79,56) und damit diese Art von Vorverarbeitung zu deutlich besseren Parsing-Ergebnissen führt als die Verwendung des „caseless“ Modells. Ein Überblick findet sich in Tabelle 3:

	F Originalschreibweise		F bestes Baseline		F Kleinschreibung		F Großschreibung	
Parsermodell	labeled	unlabeled	labeled	unlabeled	labeled	unlabeled	labeled	unlabeled
factored	76.29	80.32	75,90	80,18	72,63	77,68	35,20	48,18
pcfg_case	76.34	80.16	75,99	79,98	74,77	79,34	29,17	43,68
pcfg_caseless	76.22	80.30	76,20	80,30	76,20	80,30	76,11	80,24
shift-reduce	76.05	79.82	75,74	79,54	72,20	77,16	42,77	54,46
shift-reduce with beam search	76.80	80.90	78,05	81,86	72,08	77,10	42,93	54,92
relational neural network	78.24	82.20	77,76	81,80	76,87	81,34	24,94	40,32
dependency neural network	79.56	83.06	79,18	82,80	77,70	81,68	40,70	51,96

Tabelle 3: Vergleich der Parsermodelle bei unterschiedlicher Groß- und Kleinschreibung

Audio

Zur Vorbereitung der Audioverarbeitung mussten darüber hinaus die Untertitel von allem Text befreit werden, der nicht gesprochen wird. Wesentliche Punkte sind dabei Sprecherangaben („Reporter:“) und Angaben über den nicht-gesprochenen Ton („[Doorbell rings.]“ oder „[Applause]“). Mittels einer Frequenzliste wurde ein Filter erstellt, der ca. 95 % des nicht-gesprochenen Texts entfernt.

Im nächsten Schritt wurde mit Forced Alignment Software (in diesem Fall *Gentle*) versucht, für jedes Wort in den Untertiteln die genaue Position in der Audiospur des Videos zu ermitteln. Die Software selbst gibt an, etwas über 91 % der Wörter zu alignieren, aber Stichproben zeigen, dass auch bei den alignierten noch falsche Ergebnisse auftreten. Die genaue Größenordnung des Fehlers muss noch ermittelt werden.

Video

Schließlich wurde mittels Computer-Vision-Software die visuelle Ebene auf verschiedene Merkmale hin annotiert, die für die Erforschung multimodaler Kommunikation relevant sind. Der Computer versucht hier, automatisch zu erkennen, ob eine Person auf dem Bild zu sehen ist, ob diese der Sprecher bzw. die Sprecherin ist, ob die Person ihre Hände bewegt und ob bestimmte high-level-Gesten (in diesem Fall als Test sogenannte „timeline gestures“) zu sehen sind (Turchyn et al. 2018). Erste Tests zeigen, dass das System auf OpenCV-Basis eine sehr gute Präzision jenseits der 90% für die Personenerkennung schafft, aber leider bei den Handbewegungen nur eine Präzision im Bereich von ca. 33 % erreicht. Es ist also im Moment immer ein nachgeschalteter manueller Analyseschritt nötig. Aktuell laufen Experimente, die Erkennung mittels OpenPose zu verbessern.

Abfragemöglichkeiten

Alle Daten wurden in *CQPweb* (Hardie 2012), einer korpuslinguistischen Abfrageplattform mit großem Funktionsumfang, gespeichert und können so effizient und komfortabel abgefragt werden. Es wird gezeigt, wie mit einer Abfrage sowohl linguistische als visuelle Parameter abgefragt werden können, so dass man sofort die jeweils passenden Stellen im Video angezeigt bekommt.

Weiterhin werden Abfragemöglichkeiten über ein Geoinformationssystem (GIS) in Verbindung mit linguistischer Analyse (z.B. für die kulturgeographische Forschung) sowie die Suchmaschinen der UCLA demonstriert.

Um die oben erwähnte manuelle Analyse zu beschleunigen, wurde im Rahmen des *Google Summer of Code* 2018 die Version 2 des *Red Hen Rapid Annotator* entwickelt, mit dem komfortabel und schnell große Mengen an Videoschnipseln klassifiziert werden können. Im Vortrag wird dieser ebenfalls kurz demonstriert.

Anwendungen

Abschließend wird ein kurzer Überblick über laufende und abgeschlossene Projekte mit der Infrastruktur gegeben, um zu zeigen, welch vielfältige Fragestellungen bereits heute bearbeitet werden können.

Bibliographie

Chen, Danqi / Manning, Christopher D. (2014): *A Fast and Accurate Dependency Parser using Neural Networks*, in: **Moschitti, Alessandro / Pang, Bo / Daelemans, Walter (eds.):** *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. Doha: Association for Computational Linguistics, 740-750. http://aclweb.org/anthology/D14-1082 [Letzter Zugriff 15. Januar 2018]

Hardie, Andrew (2012): *CQPweb – Combining Power, Flexibility and Usability in a Corpus Analysis Tool*, in: International Journal of Corpus Linguistics, 17.3, 380-409

Manning, Christopher D. / Surdeanu, Mihai / Bauer, John / Finkel, Jenny Rose / Bethard, Steven / McClosky, David (2014): *The Stanford CoreNLP Natural Language Processing Toolkit*, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14). Baltimore, MD:

Association for Computational Linguistics, 55-60. http://aclweb.org/anthology/P14-5010.pdf [Letzter Zugriff 15. Januar 2018]

Steen, Francis F. / Turner, Mark (2013): *Multimodal Construction Grammar*, in: **Borkent, Michael / Dancygier, Barbara / Hinnell, Jennifer (eds.):** *Language and the Creative Mind*. Stanford, CA: CSLI Publications, 255-274

Turchyn, Sergiy / Olza Moreno, Inés / Pagán Cánovas, Cristóbal / Steen, Francis / Turner, Mark / Valenzuela, Javier / Ray, Soumya (2018): *Gesture Annotation with a Visual Search Engine for Multimodal Communication Research*, in: IAAI-18, article 72.

Uhrig, Peter (2018): *NewsScape and the Distributed Little Red Hen Lab – A digital infrastructure for the large-scale analysis of TV broadcasts*, in: **Anne-Julia Zwierlein / Jochen Petzold / Katharina Böhm / Martin Decker (eds.):** *Anglistentag 2018 in Regensburg: Proceedings. Proceedings of the Conference of the German Association of University Teachers of English*. Trier: Wissenschaftlicher Verlag Trier.