berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

Kollaborative Erstellung eines annotierten Korpus als Grundlage für die Anwendung statistischer Ansätze der automatischen Sprachverarbeitung auf internetbasierte Kommunikation

Kay-Michael Würzner | Lothar Lemnitzer | Bryan Jurish | Alexander Geyken

Webkorpora in Computerlinguistik und Sprachforschung IDS Mannheim, 28. September 2012

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

Gliederung

1. Einführung
2. Statistische Tokenisierung
3. Kollaborativ erstellte Goldstandards

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

Das große Ganze

- internetbasierte Kommunikation
  - (Mikro-)Blogs
  - Foren
  - **Chats**
- linguistische Annotation auf Wortebene
  - **Tokenisierung**
  - Lemmatisierung
  - PoS-Tagging
- Suchmaschinenindizierung
- tiefere Annotation (i.e. Syntax)

Tokenisierung – Begriff

- Unterteilung von Fließtext in Wörter (bzw. *Token*) und Sätze
- (Vor-)Klassifizierung der Token (u.a. zur Beschleunigung der morphologischen Analyse)
- Definition:

Gegeben eine Zeichenkette $w = w_1 \dots w_n$ $t : \mathbb{N} \to \mathbb{B} \times \mathbb{B}$ ist eine Funktion, so dass

$$t(i) = \begin{cases} (true, true) & \text{if } \texttt{bot}@w_i \wedge \texttt{bos}@w_i, \\ (true, false) & \text{if } \texttt{bot}@w_i, \\ (false, false) & \text{andernfalls.} \end{cases}$$

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

Herausforderungen

- Abkürzungen
- Zahlen
- Sonderzeichen
- Fremdalphabete
- Normalisierung der Silbentrennung

Regelbasierte Ansätze

- Ursprung in der *lexikalischen Analyse* von Programmiersprachen (vgl. Aho, Sethi und Ullman 1986)
- Konstruktion eines deterministischen, endlichen Automaten
- zumeist mit Hilfe eines sog. Scanner-Generators (`flex`, `re2c` etc.)
  - Patterndefinition per regulärem Ausdruck
  - resultierender Automat „hart kodiert“
  - `switch-case`-Konstrukte
- schnelle Laufzeit (mehrere MB pro Sekunde)

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

Grenzen regelbasierter Ansätze 1: Ambiguitäten

- Ambiguität einiger Zeichen bzgl.
  - `bos/eos`
  - `bot/eot`
  - Kategorie
- z.B.: '.', ':' , '', '/', `[IVXLDCM]`

Beispiele für Ambiguitäten

häufigste Fehlerquelle: „.“

Nach einer Schätzung des Industrieministeriums sind es mehr als 800.

„Österreich wurde alleingelassen in Europa“, beschwerte sich SPÖ-Zentralsekretär Josef Cap.

Satzende nach „:“

FR: Auf die Wahlerfolge der rechtsradikalen Parteien . . .

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

Beispiele für Ambiguitäten (fortges.)

Einzelfehler Tokenisierung

Kaiser's-Netz → Kaiser 's-Netz

mm. → mm. [ORD]

CDU/CSU → CDU / CSU

Jeanne d'Arc → Jeanne d' Arc

(Verwaltungs-)Personal → (Verwaltungs-) Personal

Grenzen regelbasierter Ansätze 2: resultierende Kodegröße

- komplexe Tokendefinitionen (mit Überlappungen)
- Kontextdefinitionen
- verschiedene „Zustände“
- *Lookahead, -back*

DWDS-Tokenizer: > 300 MB C++-Kode

mehrere Stunden Kompilierzeit

Hauptspeicherbedarf > 2 GB

Statistische Ansätze

- Nutzung überzufälliger Häufigkeiten zur Entscheidungsfindung
- überwachte Verfahren
  - korrekt annotiertes bzw. kategorisiertes Material zum Training
  - Laboreiro et al. (2010): Tokenizer für Twitter
- unüberwachte Verfahren
  - Training auf Rohdaten
  - basiert auf „sicheren“ Fällen
  - Kiss und Strunk (2006): Satzendeerkennung mit Hilfe von Kookkurrenzen

Ein überwachter Ansatz

- PoS-Tagger `moot` (Jurish 2003)
  - HMM-basiert (2. Ordnung)
  - satzweise Verarbeitung
  - Viterbi-Optimierung:

$$\tau(w_{i\ldots n}) = \arg \max_{t_{1\ldots n} \in T^n} P(t_{1\ldots n}|w_{1\ldots n})$$

$$P(t_{1\ldots n}|w_{1\ldots n}) = \prod_{i=1}^{n} P(t_i|w_{1\ldots i-1}, t_{1\ldots i-1}) P(w_i|w_{1\ldots i-1}, t_{1\ldots i})$$

- Training und Evaluierung mit Tiger

Trainingsphase

- Klassifizierung der Token bzgl.
  - Klasse $K = \{alpha, alpha\text{-}stopword, numeric, eos, dot, comma, quote, other\}$
  - Schreibung $S = \{upper, lower, caps, *\}$
  - Länge $L = \{1, \leq 3, \leq 5, long\}$
  - Abkürzung $A = \{known, unknown\}$
- Klassifizierung des Token-Status' bzgl.
  - Tokenanfang `bot` $\in \mathbb{B}$
  - Satzanfang `bos` $\in \mathbb{B}$

Trainingsphase (fortges.)

- Text $T \in K \times S \times L \times A$ (minus „unmögliche“ Belegungen)

  $\rightarrow$ *Verdichtung* des Lexikons

- Tag $\in T \times \mathbb{B} \times \mathbb{B}$ (minus „unmögliche“ Belegungen)

  $\rightarrow$ Tagset!

- Zuweisung des korrekten Tags

- Berechnung des HMMs samt Gewichten und Glättung per Interpolation

Trainingsphase – Beispiel

Das	cls:alpha−sw_cas:upper_abbr:uk_len:3_bos:1_bot:1
Unternehmen	cls:alpha_cas:upper_abbr:uk_len:long_bos:0_bot:1
verkauft	cls:alpha_cas:lower_abbr:uk_len:long_bos:0_bot:1
er	cls:alpha−sw_cas:lower_abbr:uk_len:3_bos:0_bot:1
1984	cls:num_cas:∗_abbr:uk_len:5_bos:0_bot:1
fuer	cls:alpha−sw_cas:lower_abbr:uk_len:3_bos:0_bot:1
2,5	cls:num_cas:∗_abbr:uk_len:3_bos:0_bot:1
Milliarden	cls:alpha_cas:upper_abbr:uk_len:long_bos:0_bot:1
.	cls:dot_cas:∗_abbr:uk_len:1_bos:0_bot:1

Evaluierung

- 10malige Kreuzvalidierung (90:10) auf Tiger
- Mittelwerte

		händisch		automatisch	
	+ Items	892 710	(100,0%)	891 110	(100,0%)
+eos	– Match	890 032	(99,7%)	889 328	(99,8%)
	– NoMatch	2 678	(0,3%)	1 782	(0,2%)

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

Dortmunder Chatkorpus

- umfassende Sammlung von Mitschnitten verschiedener Chat-Anwendungen
- Zusammenstellung an der TU Dortmund unter der Leitung von Angelika Storrer und Michael Beißwenger
- Inhalte:
  - Chats im Hochschulkontext (E-Learning, Online-Zusammenarbeit, kollektive Experten-Interviews)
  - Beratung und Support (z.B. BAFÖG-Beratung)
  - Chat-Events im Medienkontext (z.B. VIPs und sendungsbegleitende Diskussionen)

Dortmunder Chatkorpus (fortges.)

- noch Inhalte:
  - „Plauder“-Chats im Freizeitbereich (Interessensgruppen z.B. Degu-Freundinnen)
- Basiskorpus mit 140 000 Beiträgen (ca. 1 Million Token)
- frei verfügbares Releasekorpus (ca. 500 000 Token)

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

Eigenschaften von Chatsprache – explorativ

- nichtkontinuierlicher „Gesprächsverlauf“
- nichtstandardkonforme Orthographie
  - kaum Interpunktion
  - „. . . “ als Allzweckinterpunktionszeichen
  - stark reduzierte Großschreibung
  - zuällig eingestreute Selbstkorrekturen
- häufig elliptische Konstruktionen
- Gesten in Form von Emoticons, ASCII-Art und verkürzten finiten Verbformen (z.B. „*läster“)

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

Tokenisierung von Chatsprache

- Erstellung eines hand-tokenisierten Testkorpus' (ca. 40 000 Token, 12 unterschiedlich lange Chatprotokolle)
- opportunistische Zusammenstellung ($1:1$, $1:n$, $m:n$)
- Auflösung nicht-kontinuierlicher Gesprächsverläufe anhand von Äußerungs- und Sprecher-ID
- Trainierung eines Modells auf Basis der Chat-Protokolle
- Vergleich Tiger- vs. Chat-Modell

Tokenisierung – Evaluierung

		händisch		automatisch	
Tiger	+ Items	39 747	(100,0%)	37 271	(100,0%)
	– Match	36 535	(91,9%)	36 535	(98,0%)
	– NoMatch	3 212	(8,1%)	736	(2,0%)
Chat	+ Items	39 747	(100,0%)	39 158	(100,0%)
	– Match	38 515	(96,9%)	38 515	(98,4%)
	– NoMatch	1 232	(3,1%)	643	(1,6%)

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

Kollaborativ Erstellte Trainingskorpora

- verschiedene Projekte bearbeiten IBK-Daten (Empirikom)
- trotz unterschiedlicher Datengrundlagen ähnliche Phänomene
- händische Materialerstellung aufwendig (und teuer)

→ Arbeitsteilung liegt nahe

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

Agenda

- Datengrundlage
  - ausreichende Größe
  - Abdeckung aller benötigten Textsorten
  - repräsentativ und ausgewogen (DeRiK?)
- Annotationsebenen
  - mehr als PoS-Tagging und Lemmatisierung?
  - Mehrwortebene (i.e. *Chunking*)?
  - Zeichenebene?

Agenda (fortges.)

- Annotationsformat
  - bis jetzt nur Text
  - XML? (Stand-off vs. inline; Paula?)
  - geeignete Annotationswerkzeuge?
- Koordination und Qualitätskontrolle

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Universität Potsdam

Danke für Ihre Aufmerksamkeit!

Außerdem Dank an Michael Beißwenger und Gabriella Pein

Eigenschaften von Chatsprache – quantitativ

	Chatkorpus	Tiger
∅ Tokenlänge in Buchstaben	4,5	5,5
∅ Satzlänge in Wörtern	18,1	17,6
Anteil rein alphabetischer Token in Prozent	82,8	85,0
Anteil falschgeschriebene Token in Prozent	19,6	13,6
Anteil großgeschriebene Token in Prozent	15,5	31,5