Mathe III: Statistische Methoden

Teil 2

Garance PARIS

Sommersemester 2010

Sprache und Statistik in der Computerlinguistik

7.–10. Juni 2010

Natürliche (menschliche) Sprache

Ein Zitat:

. . . language is **a biological system**, *and biological systems typically are* ***"messy"***, *intricate,* **the result of evolution***ary "tinkering", and* **shaped by accidental circumstances** *and by physical conditions that hold of complex systems. . .*

(Chomsky, The minimalist program)

Sprich: "Echte" Sprache ist chaotisch und "durcheinander" (Ambiguität, Fehler, Redundanz,. . .)

Natürliche (menschliche) Sprache

Sprech- und Grammatikfehler

- *Die Poxen zum Backen* (Die Boxen zum Packen)
- *Das Verfassen der Kinderbücher und der Reiseberichte haben dem Autor viel Ruhm eingebracht*
  (Das Verfassen. . . hat. . .)
- Reparatur:
  . . . bei mir ist die Woche davor schlecht, ***also****, die Woche nach Pfingsten****, und*** *die erste Maiwoche,* ***also,*** *alles andere* ***wäre stünde*** *zur Disposition, dann würde ich mal sagen, daß wir den ersten Termin auf Montag, den neunten Mai legen. . .*

  (aus dem Verbmobil-Korpus, spontan-sprachliche Terminabsprache Deutsch-Englisch-Japanisch)

Natürliche (menschliche) Sprache

Ambiguität (lexikalisch, syntaktisch und anaphorisch)

Früher stellten die Frauen der Inseln am Wochenende Kopftücher mit Blumenmotiven her, die ihre Männer an den folgenden Montagen auf dem Markt im Zentrum der Hauptinsel verkauften.
(Bsp. von H. Uszkoreit)

Wieviele Lesarten besitzt dieser Satz?
$2 \times 2 \times 2 \times 3 \times 3 \times 2 \times 4 \times 2 \times 4 \times 2 \times 2 \times 7 \times 2 = 258.048!$

- *Früher*: Adverb oder Komparativ von früh? (2)
- *stellten*: Präteritum und Konjunktiv? (2)
- *Frauen*: Subjekt oder Objekt des Satzes? (2)
- *am Wochenende*: kann *Insel*, *Frauen* oder das Verb modifizieren (3)

Natürliche (menschliche) Sprache

Bsp. zu natürlichsprachlicher Ambiguität, 2. Teil

- *mit Blumenmotiven*:
  Instrument der Herstellung oder *gemeinsam mit*? (3)
- *her*: Partikel oder direktionale Bedeutung? (2)
- *die. . . verkauften*:
  kann jedes Plural-Substantiv modifizieren (4)
- Subjekt des Relativsatzes: *die* oder *ihre Männer* (2)
- *ihre*: kann auf jedes der Substantive referieren (4)
- *Montagen*: auch Nominalisierung von *montieren* (2)
- *der Hauptinsel*: Genitiv oder Dativ? (2)
- PPs im Relativsatz können sich in sieben Kombinationen mit NPs oder Verb verbinden (7)
- *verkauften*: Präteritum oder Konjunktiv? (2)

Kompetenz und Performanz

Zwei Begriffe, die von Chomsky festgehalten wurden:

Kompetenz

- Potenzielle, idealistische (angeborene) Fähigkeit zur Sprache bzw. Wissen um die Sprache
- Endliche Menge von Sprachregeln, die Sprecher verinnerlicht haben und die zum Verstehen und Produzieren von Sprache dienen
- Die Kompetenz beschreibt die wohlgeformten Äußerungen einer Sprache, aber man kann sie nicht direkt beobachten

Performanz

- Anwendung der zur Kompetenz gehörenden Regeln
- Tatsächlich vorkommende Äußerungen
- Zu beobachtendes Verhalten

Zwei Ansätze (Charles Fillmore)

The Armchair Linguist

He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting "Wow, what a neat fact", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like.

Zwei Ansätze (Charles Fillmore)

The Armchair Linguist

He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting "Wow, what a neat fact", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like.

The Corpus Linguist

He has all the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first words of a sentence versus as the second word of a sentence.

Armchair Linguistics

- Modellierung durch theoretische Überlegung
- Erstellung regelbasierter Modelle
- Gesucht werden Regeln,
  - die alle Fälle eines Phänomens erfassen, aber nicht übergenerieren
  - die einfach genug sind, um von einem Computer berechnet zu werden
    (kein Rückgriff auf Weltwissen usw.)
- Erfolgreich für Morphologie, Grammatiken (Grammatiktheorie), formale semantische Analyse
- Vorteile
  - Erlaubt Modellierung komplexer Phänomene ("tiefe" Analyse)
  - Negative Evidenz: Was **nicht** möglich ist
  - Ergebnis für Menschen verständlich
  - Bietet oft eine Erklärung des Phänomens an

Bsp.: Woran erkenne ich ein Adjektiv?

Wir suchen ein Regelsystem, das für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch ist

Bsp.: Woran erkenne ich ein Adjektiv?

Wir suchen ein Regelsystem, das für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch ist

1. Nächstes Wort kapitalisiert: Adj
   Sonst: NAdj

Bsp.: Woran erkenne ich ein Adjektiv?

Wir suchen ein Regelsystem, das für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch ist

1. Nächstes Wort kapitalisiert: Adj
   Sonst: NAdj
2. Nächstes Wort kapitalisiert und Wort kein Artikel: Adj
   Sonst NAdj

Bsp.: Woran erkenne ich ein Adjektiv?

Wir suchen ein Regelsystem, das für (möglichst) alle Adjektive wahr und für alle anderen Wörter falsch ist

1. Nächstes Wort kapitalisiert: Adj
   Sonst: NAdj
2. Nächstes Wort kapitalisiert und Wort kein Artikel: Adj
   Sonst NAdj
3. Nächstes Wort kapitalisiert und Wort kein Artikel und vorheriges Wort Artikel: Adj
   Sonst NAdj

Armchair Linguistics

Bsp.: Woran erkenne ich ein Adjektiv? (2. Teil)

*Ich möchte Ihnen für den Bericht über den **siebenten** Bericht über **staatliche** Beihilfen in der **europäischen** Union danken.*

Es ist schwer, korrekte und vollständige Regeln zu schreiben

- Regel 2 ist zu liberal (möchte = Adj)
- Regel 3 ist zu streng (staatliche = NAdj)
- Das System trifft eine harte Entscheidung für jede Instanz
- Keine Möglichkeit, über "Konfidenz" zu sprechen

Armchair Linguistics

Nachteile regelbasierter Systeme

- Nicht geeignet für stetige Phänomene (Kontinuen)
- Unpraktisch, um mit Ambiguität umzugehen
- Häufig präskriptiv statt deskriptiv
- Mangel an Robustheit: Schon bei kleinen Fehlern in der Eingabe bricht die Analyse ab
- Objektivität?
- Hand-Arbeit: Hoher Aufwand
  Die "English Resource Grammar" (ERG) wird seit Mitte der 90er Jahre in mehreren großen CL-Projekten entwickelt, aber es wird noch daran gearbeitet!

Die Alternative

Korpuslinguistik und Statistik

- Daten-orientierte Untersuchungen
  - Modellierung durch Sichtung von Beispielen
  - Typischerweise statistische Modelle
- Erkennung ähnlicher Muster und Regelmäßigkeiten in den Daten
- Vorteile
  - Auf Grund von Daten trainiert: Weniger Handarbeit (Einsatz maschineller Lernverfahren)
  - Bestimmung der wahrscheinlichsten Lesart
  - Robust: Können mit fehlerhafter oder unbekannter Eingabe umgehen
  - Da Alternativen gewichtet, können Modelle Übergenerierung erlauben, um Robustheit zu erreichen
  - Zugriff auf in den Daten implizites Weltwissen
  - Schnelle Modellierung neuer Domänen, Sprachen, usw.

Korpuslinguistik und Statistik

- Nachteile
  - Modelle oft nur approximativ richtig
  - Schwierige Probleme können oft nicht zuverlässig modelliert werden
  - Modelle für Menschen schwierig zu verstehen und abzuändern
  - Rein descriptiv, keine Erklärung
  - Abhängigkeit von den Daten
- Erfolgreich für:
  - das Entdecken von Neologismen, Datierung von Texten
  - Wortartenanalyse, “grobe” semantische Analyse
  - Automatische syntaktische Analyse

Korpuslinguistik und Statistik

Einige Beispiele

- Lexikalische Präferenzen
  - Wortkategorie: *bank* = Substantiv 85 %, Verb 15 %
  - Bedeutung: *bank* (river) = 22 %, *bank* (money) = 78 %
- Syntax:
  - realized + NP = 20 %
  - realized + S = 65 %
  - realized + other = 15 %
- Anaphern: *He* bezieht sich auf Englisch in 63 % der Fälle auf das Subjekt des vorigen Satzes
- Textanalyse: Autor X verwendet das Wort *bezüglich* “signifikant” öfter als Autor Y

Kleine Methodengeschichte der Computerlinguistik

- 1950er-1980er: Theoretische Linguistik
  - Linguistische Grundlagenarbeit (Grammatiktheorien)
  - Weniger Fokus auf praktische Anwendung
- Seit 1990: Verwaltung riesiger Datenmengen wird als zentrale Aufgabe der CL erkannt
  - Maschinelles Lernen als zentrale Methode
  - Hoffnung:
    "Jedes Problem lässt sich durch genügend Daten lösen"
    (Heute zum Teil enttäuscht)

"Every time I fire a linguist, the performance of our speech recognition system goes up." (F. Jelinek, 1988)

Parallel mit Nativismus vs. Empirie

Welche Rolle spielt Spracherfahrung beim Sprachenlernen?

- Nativismus: Sprache ist sehr komplex, daher muss die Fähigkeit dazu und deren Grundprinzipien beim Menschen angeboren sein

  (Vgl. Chomsky's *Principles and Parameters*:
  - Sowohl Prinzipien als auch Parameter sind Sprachuniversalien
  - Menschen kennen die Prinzipien von Geburt an, z. B. dass alle Sätze ein Subjekt haben, auch wenn es in manchen Sprachen overt (=sichtbar) weggelassen werden kann
  - Spracherwerb besteht darin, die Parameter für die eigene Muttersprache zu setzen: SVO oder OVS? usw.)
- Empirizismus: Sprachliches Wissen erwerben Kinder ausschließlich durch das Hören der Sprache ihrer Eltern

Praktisch gesehen...

Wie können wir Text-Korpora nützen?

- Ausbeuten reines Texts
  - Durch Unix-Tools: Besonders effizient durch C-Implementation und die Möglichkeit, Befehlsketten zu bauen
  - Mit geeigneten Programmiersprachen, etwa sed, awk, perl, python, java, usw.
- Aufbereitung von Texten (durch Tagger, Chunker, Parser, usw.)
- Auslesen hinzugefügter linguistischen Information (mit spezifischen Tools, z. B. Konkordanz-Programmen, oder mit allgemeinen Programmiersprachen)

Word Counts

- Word Tokens: **Gesamtzahl** Wörter im Korpus

  $Peters_1$ $Vater_2$ ist_3 $Koch_4$ $.5$ $Peters_6$ $Mutter_7$ ist_8 $Köchin_9$ $.10$

- Word Types: Anzahl **verschiedener** Wörter im Korpus

  $Peters_1$ $Vater_2$ ist_3 $Koch_4$ $.5$ $Peters_1$ $Mutter_6$ ist_3 $Köchin_7$ $.5$

  Hier muss aber entschieden werden, was als gleich zählt:

  - *Peter* und *Peters*? *bin, bist, ist,. . .*? *Koch* und *Köchin*? *Vater* und *Mutter*?
  - Homographen verschiedener Wortkategorien?
    to saw_V *the wood/sharpen the* saw_N
    das $schnelle_{ADJ}$ *Auto/der Zug fährt* $schnell_{ADV}$
    (Voraussetzung, das Korpus ist getaggt)
  - Homographen mit verschiedener Bedeutung?
    saw the wood/saw the film

- Das Type/Token Ratio: Kann zur Charakterisierung von Texten, Genres, Autoren, usw. dienen

Praktisch gesehen...

Muster-basierte Korpus-Auswertung

Anfragen auf Basis von Zeichenfolgen und regulären Ausdrücken, z. B.:

- `[word="ziemlich"][pos="ADJA"]*[pos="NN"];`

  (pos = Part-of-Speech)

  ziemlich wichtige Frage
  ziemlich kunterbuntes Bild
  ziemlich rustikales Vergnügen

- `[lemma="Hund"][pos!="$.*"]*[pos="NN"];`

  (*Hund...Mutter*, mit beliebig vielen Wörtern dazwischen, aber kein Interpunktionszeichen)

  Hund für ihre Mutter
  Hunden und Maschinengewehren
  Hunde und verschnupfte Katzen
  Hunde an die Leine

Praktisch gesehen...

Korpus-Auswertung: n-Gramme

- Wahrscheinlichkeit eines Wortes in Abhängigkeit des vorhergehenden Kontexts:
  $p(w_n) = p(w_n|w_1, \ldots, w_{n-1})$
  Bsp.: *toter Fisch* ist wahrscheinlicher als *toter Tisch*
- Wunsch: Kontext so groß wie möglich
- Einschränkungen:
  - Anzahl der zu estimierenden Parameter (Wahrscheinlichkeiten)
  - Modell unzuverlässig für ein anderes Korpus
- Markov-Annahme: Nur der unmittelbar vorhergehende Kontext hat einen Einfluss
  Bigramme (2nd order Markov),
  Trigramme (3rd order Markov)

Sprachresourcen

Überblick

- Korpora:
  Ein Korpus (n.!) ist eine endliche Sammlung von konkreten sprachlichen Äußerungen, die als Grundlage für sprachwissenschaftliche Untersuchungen dienen
  (Lexikon der Sprachwissenschaft)
- Wörterbücher, Lexika, Thesauri und Enzyklopädien
- Semantische Netze und Wissensrepräsentation (Ontologien)

Es gibt inzwischen Korpora für (fast) jede linguistische Ebene—fürs Englische!

Überblick über Sprachressourcen

Rohe Korpora

- Sehr große Korpora
  - Korpora aus dem Internet: Mehrere G Wörter
  - Das Internet selbst
- Einsatz in der Lexikographie:
  - Manuelle Sichtung der Beispiele (Konkordanz)
  - Bestimmung von Wortbedeutungen
  - Suche nach Neologismen und Kollokationen
- Probleme: Text von Formattierung trennen, Sauberkeit der Daten, Erkennung der Sprache. . .

Eine Konkordanz

hängen , Packpferde mit Brennholz ; Frauen	backen	Brot , Kinder hüten Ziegen . Von Zeit zu Zeit
unmusikalisch . Aber sie kann Pfannkuchen Brot ,	backen	Nun folgt die konkrete Utopie (oder was m
Bei 170 Grad , Gas : Stufe 3 etwa 1 1/4 Std.	backen	. Vor dem Herausnehmen erkalten lassen . R
Leute an . Laßt uns anfangen , ich muß Brot “	backen	, meinte er unwirsch und genehmigt sich un
kann doch nicht jeder seine eigenen Brötchen .	backen	, mahnte Scherf . Dann wieder Fragen : Ob
e , und die zieht er formvollendet durch : Wir	backen	einen guten Kurzfilm . An der Idee blieb auc
ssen . In heißem Backfett kleine Pfannkuchen	backen	und mit saurer Sahne und Kaviar servieren .
zu besticken , Kaffee zu kochen und Kekse zu	backen	, um so ihrer Verpflichtung gegenüber dem
. Im Moment aber muß er ganz kleine Brötchen	backen	. Der Grüne sieht sich einer erdrückenden sc
, 1/2 Stunde ziehen lassen , dann goldbraun	backen	. Mit Erdbeeren garnieren . Alle Rezepte aus
Halloween höhlen sie einen Kürbis aus und “	backen	Pumpkin-Pie . Die Prices sind eine durchsch
ade oder Quark . Schwaben südlich der Donau	backen	Brot , wie die riesigen Knauzawecka , noch
schwimmen gehen , nachtwandern , Stockbrot	backen	, die Bauern besuchen , basteln , spielen . Bei

Überblick über Sprachressourcen

Korpora mit Wortarten

- Standardkorpora:
  - British National Corpus (BNC), 100M Worte
  - American National Corpus (ANC), 22M Worte
  - Huge German Corpus (HGC), 200M Worte
- Einsatz: Training von Taggern
- Formatierung:
  - Ad-hoc Format
  - SGML Mark-up (British National Corpus)
  - Interlinear format (hier Wort_PoS_Lemma):
    John_PN_john left_VBP_leave ._PUNC_period
  - Spalten (Susanne, 1. Spalte: Satz- und Wort-Id)

A12:0210	John	john	PN
A12:0211	left	leave	VBN
A12:0212	.	Period	PUNC
  - Heute: meist XML (Vorteil: Allgemeine Tools)

Beispiel aus dem BNC (SGML)

- Veraltet, "Mutter" von XML
- Anfang- und Endtags optional
  (wegen Ersparnis an Speicherplatz)

```
<s n=0001>
 <w NN1>INTRODUCTION
</head>
<p>
<s n=0002>
 <w AT0>The <w AJ0>extensive <w NN1>upland <w NN2-VVZ>landscapes <w PRF>of
 <w AT0>the <w NP0>UK<c PUN>, <w CJC>and <w AT0>the <w AJ0>varied <w CJC>and
 <w AJ0>rich <w NN1>wildlife <w PNP>they <w VVB>support<c PUN>, <w VBB>are <w AT0>the
 <w NN1>product <w PRF>of <w NN2>centuries <w PRF>of <w AV0>predominantly
 <w AJ0>pastoral <w AJ0-NN1>agricultural <w NN1>activity<c PUN>.
<s n=0003>
 <w PRP>In <w AT0>the <w AJ0-NN1>past<c PUN>, <w AT0>the <w NN1>use <w PRF>of
 <w DT0>these <w NN2>uplands <w PRP>for <w NN0>sheep <w CJC>and <w NN1>beef
 <w NN2>cattle <w NN1-VVG>rearing <w VHZ>has <w XX0>not <w VVN>conflicted
 <w AV0>significantly <w PRP>with <w AT0>the <w NN1>need <w TO0>to <w VVI>retain
 <w NN2>habitats <w PRP>such as <w NN2>moorlands<c PUN>, <w NN1>hill
 <w NN2>grasslands<c PUN>, <w AJ0>high <w NN1>altitude <w AJ0>montane
 <w NN1>vegetation<c PUN>, <w AJ0-VVD>enclosed <w NN2>pastures <w CJC>and
 <w NN1-VVB>hay <w NN2>meadows<c PUN>, <w NN2>wetlands <w CJC>and <w AJ0>native
 <w NN2>woodlands<c PUN>, <w DTQ>which <w VVB>form <w AT0>the <w NN1>basis <w PRF>of
 <w AT0>the <w NN1>nature <w NN1>conservation <w NN1>interest <w PRF>of <w AT0>the
 <w CRD>9.68 <w CRD>million <w NN2>hectares <w PRF>of <w NN1>upland <w PRP>in
 <w AT0>the <w NP0>UK<c PUN>.
```

Überblick über Sprachressourcen

Syntax-Korpora ("Baumbanken")

- Penn Treebank: 1M Worte aus dem Wall Street Journal
- Deutsch:
  - NEGRA
    (20.000 Sätze Frankfurter Rundschau, 400K Worte)
  - TIGER
    (50.000 Sätze Frankfurter Rundschau = 1M Worte)
- Prague Dependency Treebank (Czech)
- Neuerdings auch für viele andere Sprachen:
  Chinesisch, Französische, usw.

Überblick über Sprachressourcen

NEGRA

- Als SQL Datenbank gespeichert;
  kann man in Bäume umwandeln
- Annotation:
  - PoS-tagged
  - Morphologische Annotation (60K)
  - Grammatische Funktionen
- Vorgehen:
  - Kombination aus automatischer Analyse und menschlicher Arbeit
  - Abfrage mit speziell dafür entwickelte Tools

Überblick über Sprachressourcen

Semantik-Korpora

[Peter]	Agent
gibt	
[Maria]	Recipient
[ein Buch]	Theme

- Satzteilen werden semantische Rollen zugeordnet
- Einsatz: Training semantischer Parsern
- Korpora:
  - Englisch: PropBank, auf Grundlage der Penn Treebank
  - Deutsch: SALSA, auf Grundlage von TIGER

Diskurs-Korpora

[Peter ist müde]. Grund
Deshalb [schläft er]. Folge

- Ordne Paaren von Sätzen Diskursrelationen zu:
  z. B. Begründung (weil), Zweck (damit),. . .
- Training von “Diskurs-Parsern”
- Korpora:
  DiscourseBank, auf Grundlage der Penn Treebank

Überblick über Sprachressourcen

Bilinguale Korpora

- Vergleichbare Daten:
  Crater corpora (English, French, Spanish)
- Parallel: Hansard Corpus, EUROPARL

Keine Korpora verfügbar

- Pragmatik:
  Intention der Sprecher, “was wirklich gemeint ist”
- Viele andere Sprachen, besonders für höhere Ebenen

Merkmale von Korpora

- Sprache: monolingual vs. bilingual vs. multilingual; vergleichbar vs. parallel, aligniert
- Textart, Inhalt, Genre, Domäne:
  - Spontansprache: Usenet, Wizard-of-Oz Experimente
  - Editiert: Zeitungsartikel, Romane, Fachtexte, Lyrik,...
  - Ausgewogenheit:
    homogen vs. heterogen, unbalanciert vs. balanciert
- Geschriebene Sprache vs. gesprochene Sprache
- Umfang (Tokens, Types), Zeitraum
- Format (s. oben), Text oder Binär (indexiert)
- Medium (Text, Audio, Transkripte, Video, usw.)
- Aufbereitung und Annotation
- Urheber- und Nutzungsrechte, Preis
- Standard-Referenz: Allgemeine Verfügbarkeit

Entwicklung von Sprachressourcen

- Roher Text genügt oft nicht, daher wird es ergänzt um Satzgrenzen, Wortkategorien,. . .
- Korpus-Annotation macht enthaltene linguistische Information explizit, kann aber falsch sein

Prinzipien für Korpus-Annotation (Leech, '93)

- Sowohl Annotation als originales (rohes) Korpus sollte von einander trennbar sein
- Die Annotation sollte Theorieunabhängig und neutral sein
- Die Annotationsmethode (manuell, maschinell, oder Kombination davon) sollte bekannt sein
- Die Annotationsrichtlinien sollten mit allen Details verfügbar sein

Entwicklung von Korpora: Vorverarbeitung

Korpus-Aufbereitung: Säuberung

- Im Idealfall enthält die Eingabe einen grammatischen, kohärenten Text
- Häufige Probleme:
  - Eingescannte Texte: OCR-Fehler
  - Webseiten: HTML-Markup, Navigationselemente
  - Zeitungsartikel: Insets etc.
  - Blogs/Forums:
    Unvollständige und ungrammatische Sätze

Entwicklung von Korpora: Vorverarbeitung

Korpus-Aufbereitung: Strukturierung

- Meta-Information
  - Quelle, Autor, Datum, ev. Annotator
  - Bei Dialogen: Sprecher-Details
- Audio-Korpora
  - Time-Stamps
  - Phonetische Transkription
  - Prosodie und Pausen
  - Überlappungen bei mehreren Sprechern
- Text-Korpora
  - Titeln, Untertiteln, Inhaltsverzeichnis, Fußnoten, Tabellen und Abbildungen, Stichwortverzeichnis, usw.
  - Absätze (Fließtext) oder Turns (Dialoge)
  - Fehlstarts, Reparaturen, Neuanfänge, Zögerungen

Entwicklung von Korpora: Vorverarbeitung

Korpus-Aufbereitung: Satzgrenzen

I spoke to Mr. and Mrs. Gore from Washington D.C. today.
"You remind me," she remarked, "of your mother."

- Problem: Einsätze, Abkürzungen
- Baseline:
  - Satzgrenze nach Punkt, Fragezeichen, Ausrufezeichen
  - Erreicht ca 90 % Korrektheit
- Regelbasierte Ansätze: s. Manning & Schütze
- Statistische Ansätze sind besser

Entwicklung von Korpora: Vorverarbeitung

Korpus-Aufbereitung: Tokenisierung

Frage: Was ist ein Wort?
Eine Folge alphanumerischer Zeichen,
durch Leerzeichen (allg.) oder Interpunktion getrennt

- Interpunktion von Wörtern trennen,
  wenn keine Abkürzung
- Apostrophen ersetzen? *Robert's*, *isn't* vs. *it's*, *geht's*
- Striche interpretieren: Gedankenstrich, Komposita,
  Bindestrich, Umbrüche an Zeilenenden
- Spezielle Tokens: Zahlen (2 000), "Named entities" (Namen,
  Daten, Zitate, Adressen, Telefonnummer, usw.), multi-word
  expressions

Entwicklung von Korpora

Annotation: Lemmatisierung

- Def.: Wörter mit deren Grundform versehen
- Probleme bei Homographen, s. oben
- Vereinfachungsmöglichkeiten:
  - Großschreibung eliminieren
  - Satzinterne Interpunktion löschen
  - Stemming (Affixe löschen) z. B. *operating*, *operate*, *operated*, *operates*, *operator: operat*
- Wird heutzutag nicht mehr gemacht, weil zu viele wichtige Information verloren geht

Entwicklung von Korpora

Annotation: PoS-Tagging

- Manuell oder automatisch?
- Tag Sets sind unterschiedlich groß; sie varieren in sowohl innerhalb als auch unter Kategorien in ihrer Granularität

	Brown	Penn	Claws 1–8	STTS
Größe	77/177	45	60–160	54

- Sie sind sprachspezifisch
- Manchmal enthalten sie Seltsamkeiten
  - Brown:
    VBG für Present Participles und für Gerunde
    John is purchasing apples
    The Fulton County purchasing department
  - Penn:
    TO sowohl für Präpositionen als auch vor Infinitiven
    (*I want to go to the store*)

Brown Tagset

–	dash
,	comma
:	colon
.	sentence closer (. ; ? *)
(	left paren
)	right paren
*	not, n't
ABL	pre-qualifier (quite, rather)
ABN	pre-quantifier (half, all)
ABX	pre-quantifier (both)
AP	post-determiner
AT	article (a, the, no)
BE	be
BED	were
BEDZ	was
BEG	being
BEM	am
BEN	been
BER	are, art
BEZ	is
CC	coordinating conjunction
CD	cardinal numeral
CS	subordinating conjunction
DO	do
DOD	did
DOZ	does
DT	sg. determiner (this, that)
DTI	sg. or pl. det. (some, any)
DTS	pl. determiner (these, those)
DTX	double conjunction (either)
EX	existential there
FW	foreign word
HV	have
HVD	had (past tense)
HVG	having
HVN	had (past participle)
HVZ	have, pres., 3rd p. sg.
IN	preposition
JJ	adjective
JJR	comparative adjective
JJS	semantic superl. adj. (chief, top)
JJT	superlative adjective
MD	modal auxiliary
NC	cited word
NN	singular or mass noun
NNS	plural noun
NP	proper noun
NPS	plural proper noun
NR	adverbial noun
OD	ordinal numeral
PN	nominal pronoun
PP$	determiner, possessive
PP$$	pronoun, possessive
PPL	sg. reflexive pers. pron.
PPLS	pl. reflexive pers. pron.
PPO	personal pronoun
PPS	3rd p. sg. nom. pron.
PPSS	other nominative pers. pron.
QL	qualifier (very, fairly)
QLP	post-qualifier (enough, indeed)
RB	adverb
RBR	comparative adverb
RBT	superlative adverb
RN	nominal adverb (here, indoors)
RP	particle (about, off, up)
TO	to (before infinitive)
UH	interjection
VB	verb, base form
VBD	verb, past tense
VBG	pres. part./gerund
VBN	verb, past part.
VBZ	verb, 3rd p. sg. pres.
WDT	wh- determiner
WPO	wh- pronoun, object
WPS	wh- pronoun, nom.
WQL	wh- qualifier (how)
WRB	wh- adverb

BNC Tagset, CLAWS 5 or C5

AJ0	Adjective
AJC	Comparative adjective
AJS	Superlative adjective
AT0	Article
AV0	Adverb
AVP	Adverb particle (e.g. up, off, out)
AVQ	Wh-adverb
CJC	Coordinating conjunction
CJS	Subordinating conjunction
CJT	that
CRD	Cardinal number
ORD	Ordinal numeral
DPS	Possessive determiner or pronoun
DT0	General determiner-pronoun
DTQ	Wh-determiner-pronoun
EX0	Existential there
NN0	Common noun, neutral for number
NN1	Singular common noun
NN2	Plural common noun
NP0	Proper noun
PNI	Indefinite pronoun
PNP	Personal pronoun
PNQ	Wh-pronoun
PNX	Reflexive pronoun
POS	The genitive marker 'S or '
PRF	The preposition OF
PRP	Preposition, except OF
PUL	Punctuation: left bracket
PUN	Punctuation: general
PUQ	Punctuation: quotation mark
PUR	Punctuation: right bracket
TO0	Infinitive marker TO
UNC	Foreign words
VBB	The present tense of BE, except is
VBD	The past tense of BE
VBG	The -ing form of BE
VBI	The infinitive of BE
VBN	The past participle of BE
VBZ	IS, 'S
VDB	The finite base form of DO
VDD	The past tense of DO
VDG	The -ing form of DO
VDI	The infinitive of DO
VDN	The past participle of DO
VDZ	The -s form of DO
VHB	The finite base form of HAVE
VHD	The past tense of HAVE
VHG	The -ing form of HAVE
VHI	The infinitive of HAVE
VHN	The past participle of HAVE
VHZ	The -s form of HAVE
VM0	Modal auxiliary verb
VVB	The finite base of lexical verbs
VVD	The past tense of lexical verbs
VVG	The -ing form of lexical verbs
VVI	The infinitive of lexical verbs
VVN	The past participle of lexical verbs
VVZ	The -s form of lexical verbs
XX0	NOT or N'T
ITJ	Interjection
ZZ0	Alphabetical symbols

Entwicklung von Korpora

Syntaktische Annotation

- Bäume oder Graphen?
- Welcher Grammatikformalismus?
- Was stellen die Knoten dar?
- Was stellen die Kanten dar?
- Chunker: Tool, das eine partielle Analyse mancher Konstituenten im Satz erstellt

Entwicklung von Korpora

Probleme bei der Annotation

- Wegen Ambiguität ist Annotation nicht einfach
- Handarbeit ist aufwendig, fehlerbehaftet, möglicherweise inkonsistent
  - Annotationsaufwand für ein Wort: 30 Sekunden
  - 1M Worte: 500 000 Minuten = 5 Jahre
  - Plus Aufwand fuer Qualitätssicherung
- Automatische Annotation ist nicht 100 % zuverlässig und macht systematische Fehler

Entwicklung von Korpora

Mögliche Lösungen

- Annotationsmöglichkeiten gering halten
  (z. B. kleines Tagset), um schwierige Entscheidungen aus dem Weg zu gehen
- Bei Unsicherheit mehrere Tags zuweisen
  (dem User wissen lassen, dass es Unsicherheit gab)
  z. B.: "Ambiguity Tags" im BNC
  AJ0-AV0 (Adjectiv oder Adverb), mit Präferenz für AJ0
- Automatische Annotation mit der Überprüfung durch menschliche Annotatoren kombinieren
- Bootstrapping