NLP – Eigenschaften von Text

Dr. Andreas Hotho
Dominik Benz
Beate Krause

Sommersemester 2008

ENDOWED CHAIR OF THE HERTIE FOUNDATION
Knowledge and Data Engineering
DEPARTMENT OF MATHEMATICS & COMPUTER SCIENCE

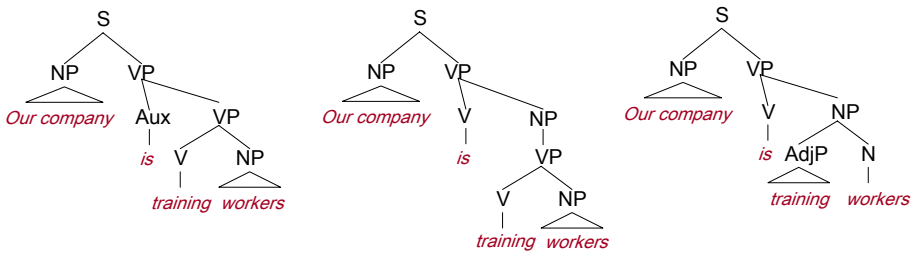
Übersicht

Inhalt - Eigenschaften von Text

- Warum NLP „schwierig“ ist
- Untersuchungsgegenstand Text
  - Resourcen, Corpora
  - Worthäufigkeiten, Zipf‘s Gesetz
  - Kollokationen, Konkordanzen
- Linguistische Grundlagen
  - Wortarten / Morphologie
    - Nomen, Pronomen
    - Begleiter, Adjektive
    - Verben
    - andere Wortarten
  - Satzstruktur
  - Semantik, Pragmatik

Warum NLP „schwierig“ ist

- Kernfrage: Welche Strukturen im Text „ergeben Sinn“?
- Viele verschiedene Zerlegungen („parses“) möglich:
  „Our company is training workers“



Unsere Firma trainiert Angestellte

(„Normaler Sinn“)

~ Unsere Firma ist, Angestellte zu trainieren

(Vgl. „Our problem is training workers“)

~ Unsere Firma ist trainierende Angestellte

(Vgl. „Those are training wheels“)

Warum NLP „schwierig“ ist

- 455 parses (!) möglich für
  „List the sales of the products produced in 1973 with the products produced in 1972“
- NLP-System müssen **sehr gut** zahlreiche Unterscheidungen bzgl. lexikalischen / strukturellen Eigenschaften (Wortsinn, Wortkategorie, syntaktische Struktur, semantische Abgrenzung) treffen können
- Symbolische NLP-Systeme haben das Problem:

- Statistischer Ansatz: automatisches Lernen der wahrscheinlichsten Eigenschaften aus Text-Korpus

Untersuchungsgegenstand Text - Resourcen

- Textsammlungen:
  - Brown Corpus (~1960), Penn Treebank Corpus (Texte aus Wallstreet Journal), …
  - „balanced corpus“: (möglichst) repräsentatives Sample der Sprache
- Bilinguale Textsammlungen (mit parallelen Texten)
  - z.B. Canadian Hansards
  - wichtig für statistische maschinelle Übersetzung
- Wörterbücher: WordNet
  - strukturierte Sammlung englischer Wörter
  - organisiert in „Synsets“ + verschiedenen Beziehungen zwischen diesen

Worthäufigkeiten

- Erste Analyse: was sind die häufigsten Wörter?
  - Bsp: Mark Twain‘s „Tom Sawyer“:

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Worthäufigkeiten (2)

- Beobachtung: sehr wenige Worte kommen sehr häufig vor, sehr viele dagegen sehr selten:

Word Frequency	Frequency of Frequency
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11–50	540
51–100	99
> 100	102

Zipf's Gesetz

- Formuliert basierend auf dem Prinzip des geringsten Auwandes:
  - Häufigkeit eines Wortes: *f*
  - „Rang“ eines Wortes: *r* (häufigstes Wort → *r* = 1, …)
  - Es existiert eine Konstante *k* so dass

    $$f \bullet r = k$$

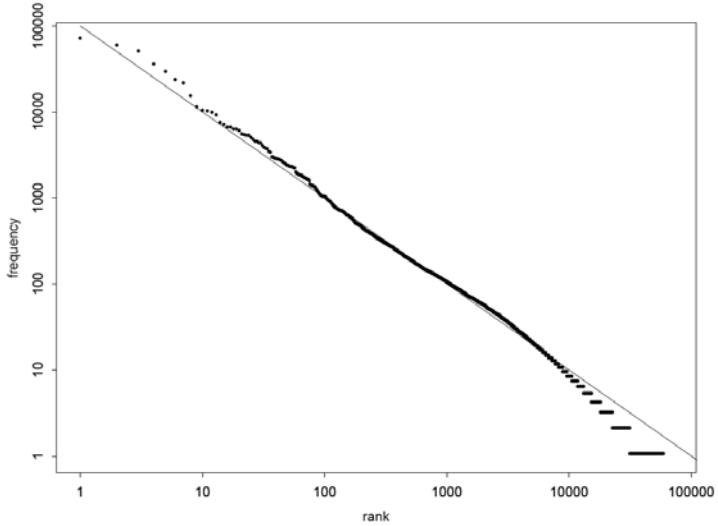
  → linearer Zusammenhang Rang / Häufigkeit auf log / log – Skala
- Verfeinerung von Mandelbrot

  $$f = P(r + \rho)^{-B}$$

  (P, B, ρ Kennzahlen für Reichhaltigkeit des Texts)

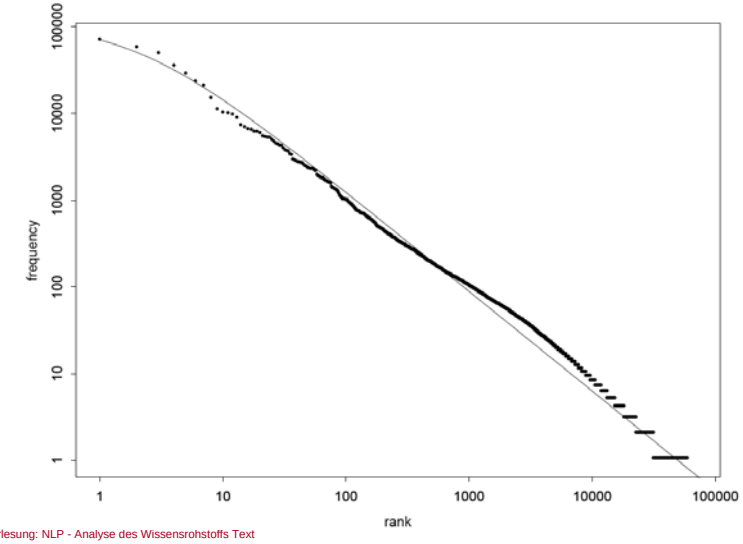
Zipf's Gesetz (2)

- Ergebnisse auf Brown Corpus:



Verfeinerung von Mandelbrot

- Ergebnisse auf Brown Corpus:



Kollokationen

- Gehäuftes gemeinsames Auftreten von Wörtern:
  - zusamengesetzte Ausdrücke (*„disk drive“*)
  - zusammengesetzte Verben (*„make up“*)
  - andere Standardausdrücke (*„bacon and eggs“*)
- Oft, aber nicht notwendigerweise spezielle / idiomatische Bedeutung
- Häufig Untersuchung von benachbarten Wörtern („Bigramme)“
  - Nachbarschaft aber nicht zwingend (*„make [something] up“*)
- Hilfreich: Filterung der Bigramme nach Wortarten

Kollokationen (2)

- Häufigste Bigramme in der New York Times:

Frequency	Word 1	Word 2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Kollokationen (3)

- Häufigste Bigramme NYT, nur Nomen/Nomen + Adkektiv/Nomen:

Frequency	Word 1	Word 2	POS pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N