

Einführung in die Linguistik

Vorverarbeitung von Texten

Vitor Fontanella

Hochschule Hannover, Abteilung Information und
Kommunikation

Folien von
Christian Wartena
genommen.

07. September 2022

Wortarten

Wortarten

- Traditionell werden Wörter in verschiedene Klassen, wie Verben, Substantive, Adjektive usw. eingeteilt.
- Es gibt nicht ein eindeutiges Kriterium für die Einteilung in Wortklassen.
- Die Klassen sind nicht in allen Sprachen gleich.
 - aber trotzdem ziemlich ähnlich.
- Kriterien sind eine Mischung aus:
 - Morphologie
 - Syntax
 - Semantik
- Jede Grammatik oder linguistische Theorie kann andere Klassen definieren
- Ältestes Kategoriensystem: Dionysios Thrax (2nd Jhdt. v.C.)
 - Im Wesentlichen bis heute in Verwendung.

POS Tagging

Ziele

- POS = Part of Speech = Wortart
- Tagging = Labeling

Ziele

- Satzstruktur baut auf Wortarten auf.
- Beziehungen zwischen Wörtern können mit Mustern von Wortarten gefunden werden.
 - Komplexe Sachverhalte können durch Verb mit Argumente (Subjekt, Objekt) gefunden werden.
- Das Gewicht von Wörter für Retrieval hängt vom Wortart ab.
- Schlagwörter, Indexterme, Fachbegriffe sind häufig Substantive oder Substantivgruppe
 - Adjektiv + Substantiv
 - Substantiv + Artikel (Genitiv) + Substantiv

Probleme

Probleme

- Wörter können zu mehreren Klassen gehören
 - Im Kontext ist immer nur eine Wortart möglich (ausgenommen strukturelle Mehrdeutigkeiten)
 - Beispiele: Fliegen – fliegen / Tagen – tagen
- Unterschiedliche Wörter können den gleichen abgeleiteten Form haben
- Wie detailliert sind die Klassen (ggf. berücksichtigt man Unterklassen)
- Unbekannte Wörter
- Sehr viele Wortformen

Probleme

Why is POS-Tagging difficult?

- Lookup the POS in a dictionary
- Many words are ambiguous (especially in English)
- Words can be used in an unusual way
- Unknown words

Example

- (1) a. I really need some **sleep**.
b. This caravan can **sleep** up to four people.
c. You should **sleep** 8 hours a day.
- (2) a. She opened the **book** to page 37 and began to read aloud.
b. I can **book** tickets for the concert next week.

Probleme

Example

- (3) a. **Where** Susy has trouble coloring inside the lines, Johnny has already mastered shading. (Sub. Conj)
- b. **Where** are you going? (Adverb)
- c. He lives within five miles of **where** he was born. (Pronoun)
- d. Finding the nymph asleep in secret **where**. (Noun)

From: <https://en.wiktionary.org/wiki/where#Adverb>

Tagsets

POS-Tagging

- POS = Part of Speech = Wortart
- POS-Tagging: automatic annotation von Texten mit Wortarten
- SMenge der mögliche Klassen wird tagset genannt
- von 10 bis weit über 100
 - Brown Corpus Tagset: 87 Tags
 - Pennsylvania Treebank Tagset: 45 Tags (Subset of previous one)
- Including subclasses
- Including inflection (e.g. noun-plural, verb-3-pers-sg-pres-tense)
- Pragmatic labels for hard cases (like “to”)
- Classes for interpunction, special signs, etc.

Penn Treebank Tagset

CC	Coordinating conjunction	PP	Personal pronoun
CD	Cardinal number	PP\$	Possessive pronoun
DT	Determiner	RB	Adverb
EX	Existential there	RBR	Adverb, comparative
FW	Foreign word	RBS	Adverb, superlative
IN	Preposition or subordinating conjunction	RP	Particle
JJ	Adjective	SYM	Symbol
JJR	Adjective, comparative	TO	to
JJS	Adjective, superlative	UH	Interjection
LS	List item marker	VB	Verb, base form
MD	Modal	VBD	Verb, past tense
NN	Noun, singular or mass	VBG	Verb, gerund or present participle
NNS	Noun, plural	VCN	Verb, past participle
NP	Proper noun, singular	VBP	Verb, non-3rd person singular present
NPS	Proper noun, plural	VBZ	Verb, 3rd person singular present
PDT	Predeterminer	WDT	Wh-determiner
POS	Possessive ending	WP	Wh-pronoun
		WP\$	Possessive wh-pronoun
		WRB	Wh-adverb

9 more for special signs and
interpunction

Stuttgart-TübingenTagset

- Nomina
 - NN normale Nomina
 - NE Eigennamen
- Adjektive
 - ADJA attributive Adjektive
 - ADJD prädikative oder adverbiale Adjektive
- Zahlen
 - CARD Kardinalzahlen
- Verben finite Formen, ohne Imperativ
 - VMFIN modale, finite Verben
 - VAFIN auxiliare, finite Verben
 - VFIN finite Verben, voll
- finite Formen im Imperativ
 - VAIMP auxiliare Verben im Imperativ
 - VVIMP andere Verben im Imperativ
- Infinitiv
 - VVINF Infinitiv, voll
 - VAINF Infinitiv, aux
 - VMINF Infinitiv, modal, Ersatzinfinitiv
 - VVIZU Infinitiv mit "zu"
- Partizip Perfekt
 - VVPP nicht flektiertes, voll
 - VMPP Partizip Perfekt, modal
 - VAPP Partizip Perfekt, aux
- Artikel
 - ART bestimmter/unbestimmter Artikel
- Pronomina Personalpronomen
 - PPER irreflexives Personalpronomen
 - PRF reflexives Personalpronomen
- Possessivpronomen
 - PPOSAT attributierendes Possessivpronomen
 - PPOSS substituierendes Possessivpronomen
- Demonstrativpronomen
 - PDAT attribuirendes Demonstrativpronomen
 - PDS substituierendes Demonstrativpronomen
- Indefinitpronomen
 - PIAT attr. Indefinitpron. ohne Determiner
 - PIDAT attr. Indefinitpron. mit Determiner
 - PIS subst. Indefinitpron.
- Relativpronomen
 - PRELAT attr. Relativpronomen
 - PRELS subst. Relativpronomen
- Interrogativpronomen
 - PWAT attr. Interrogativpronomen
 - PWS subst. Interrogativpronomen
 - PWAV adverbiales Interrogativpronomen

Verfahren

Regelbasiert

- Handgeschriebene Regel
- Gelernte Regel

Statistisch

- Wahrscheinlichkeit, dass ein Wort ein bestimmtes Tag hat.
- Wahrscheinlichkeit, dass eine Wortendung ein bestimmtes Tag hat.
- Wahrscheinlichkeit dass eine bestimmte Folge von Tags vorkommt

Ist gut gut genug?

- Unter „Laborbedingungen“ über 95% Korrekt.
- Aber:
 - Fehler in Zeichensatz, Tokenization, Satzerkennung, haben große Folgen.
 - Diskrepanz zwischen Training- und Testdaten.
 - Der eine Fehler ist der andere nicht.
 - 10 Wortarten leichter als 200!
- Sind 5% Fehler akzeptabel?
 - Automatische Übersetzung: ???
 - Text Mining: Ja!

Vorverarbeitungsschritte

