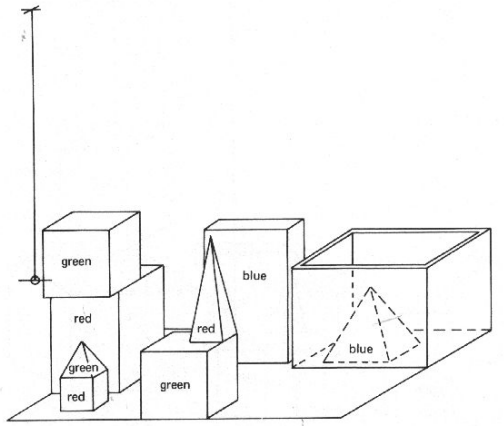
Einführung in die Computerlinguistik

WS 2008/2009

Manfred Pinkal

SHRDLU: Ein wissensbasiertes Dialogsystem



Winograds "Blocks World"

SHRDLU [2]

SHRDLU ist ein wissensbasiertes Dialogsystem.

Im Unterschied z.B. zu ELIZA verwendet es bei der Dialogführung unterschiedliche Arten von Wissen, und zwar

- Sprachliches/linguistisches Wissen
- Kontextwissen
- Weltwissen

Sprachliches Wissen in SHRDLU: Beispiele

Grammatik	Lexikon
	Morphologisches Wissen:
regelmäßige Verben bilden Präteritum auf -ed	*grasp* ist regelmäßiges Verb *put* ist unregelm. Verb mit Prät. *put*
	Syntaktisches Wissen:
In Imperativen steht das Verb an erster Stelle	*grasp* ist transitives Verb *stop* ist intransitives Verb
	Semantisches Wissen:
A+N in attributiven Konstruktionen bezeichnet Dinge, die unter A und unter N fallen	*red* bezeichnet rote Objekte *pyramid* ist Unterbegriff von *block* grasp ...

Grammatisches und lexikalisches Wissen

- Morphologische, syntaktische, semantische Regularitäten sind tendenziell in der Grammatik kodiert
- Spezielle morphologische, syntaktische, semantische Information über Einzelwörter sind im Lexikon kodiert.

- Achtung:
  - Es gibt keine scharfe Grenze zwischen systematischer grammatischer Information und wortspezifischer lexikalischer Information.
  - Unterschiedliche linguistische Theorien schlagen eine unterschiedliche Arbeitsteilung zwischen Grammatik und Lexikon vor.

Außersprachliches Wissen

- Kontextwissen:
  - Sprachlicher Kontext / Dialoggeschichte: Welches Objekt wurden zuletzt erwähnt? (*Put it into the box.*)
  - Situationskontext: Welche Objekte kommen in der Äußerungssituation vor? (*What is the pyramid supported by?*)
- Weltwissen:
  - Episodisches Wissen: Wissen über Einzelfakten
    "Es sind zwei rote Klötze auf dem Tisch."
    "Die Kiste enthält eine Pyramide"
  - Regelwissen: Wissen über Regulariten aller Art (physikalische Gesetze, Alltagswissen, ...)
    "Zwei Objekte können nicht den gleichen Platz einnehmen."
    "Ein Objekt muss eine ebene Auflagefläche besitzen, damit ein zweites stabil darauf stehen kann"

Wozu wird Wissen eingesetzt?

Wissen wird in der – menschlichen und maschinellen – Sprachverarbeitung eingesetzt, um – linguistische und extralinguistische – Strukturen unterschiedlicher Arten und Ebenen aufeinander abzubilden:

- Speech →Text
- Text → Speech
- Wortkette → Bedeutungsinformation
- Bedeutungsinformation → Handlungsplan
- Bedeutungsinformation → Wortkette
- deutscher Satz → englischer Satz

Das zentrale Problem ist die Mehrdeutigkeit (Ambiguität) auf allen Ebenen: Wie kommen wir zu einer eindeutigen Abbildung (Disambiguierung)? Hierzu ist typischerweise Kontext- und Weltwissen nötig.

Woher kommt das Wissen?

- Manuelle Grammatik- und Lexikon-Entwicklung, Erstellung von extralinguistischen Wissensbasen (Ontologien)
  - Verlässliche Information
  - Erlaubt die Modellierung komplexer struktureller Zusammenhänge
  - Extrem teuer, deshalb Abdeckungsprobleme
  - Wenig flexibel und fragil (z.B. in Bezug auf fehlerhafte Eingaben)
  - Disambiguierung ist ein großes Problem
- Automatische Akquisition von Wissen aus Sprachdaten (Korpora) mit statistischen Verfahren
  - Vergleichsweise preiswert und effizient
  - Robuste Verfahren mit hoher Abdeckung, leisten neben der Analyse auch Disambiguierung
  - Das Wissen bleibt oft implizit, im statistischen Modell
  - Approximativ korrekt, die Verlässlichkeit nimmt mit zunehmender Komplexität der linguistischen Strukturen deutlich ab
- Für bestimmte Aufgaben eignen sich Verfahren, die auf explizit kodiertes Wissen zugreifen, für andere Aufgaben statistische Verfahren, immer häufiger werden manuell kodierte und statistische Information kombiniert

Morphologie und endliche Automaten

Morphologie

- Morphologie ist der Teilbereich der Linguistik, der sich mit der internen Struktur von Wörtern befasst.
- Die wesentlichen Aufgaben der Morphologie in der Computerlinguistik sind
  - die Reduktion komplexer Wörter bzw. Wortformen auf ihre Bestandteile
  - die Identifikation von grammatischer Information, die in der Wortform kodiert ist (z.B. Kasus, Numerus, Tempus)

Teilbereiche der Morphologie

- Flexion:
  Deklination, Konjugation von Substantiven, Verben, Adjektiven, Pronomina
  frag+te+st
  ge+frag+t
- Derivation (Ableitung)
  Er+kenn+ung
- Komposition (Zusammensetzung)
  Sprach+erkennung+s+technik

Elemente der morphologischen Struktur

- Stämme, Präfixe, Suffixe (in Flexion und Derivation)
  - Flexionsmorphologie:
    frag+te+st
    ge+frag+t
  - Derivationsmorphologie:
    Er+kenn+ung

Elemente der morphologischen Struktur

- Stämme, Präfixe, Suffixe (in Flexion und Derivation)
  - Flexionsmorphologie:
    frag+te+st
    ge+frag+t
  - Derivationsmorphologie:
    Er+kenn+ung

Elemente der morphologischen Struktur

- Stämme, Präfixe, Suffixe (in Flexion und Derivation)
  - Flexionsmorphologie:
    frag+te+st
    ge+frag+t
  - Derivationsmorphologie:
    Er+kenn+ung

Elemente der morphologischen Struktur

- Grund- , Bestimmungswörter, Fugenelemente (in der Komposition)
  Sprach+erkennung+s+technik
  Sprach+erkennung+s+technik
- Bestimmungswort: *Sprach* + Grundwort: *erkennung*
- Bestimmungswort: *Spracherkennung* +Fugenelement: *s* + Grundwort: technik
- Fugenelemente sind keine Flexionssuffixe. Sie sind aufgrund der phonologischen Struktur nur partiell vorhersagbar und müssen deshalb im Lexikon gelistet werden.

Elemente der morphologischen Struktur

- Stämme, Präfixe, Suffixe (in Flexion und Derivation)
- Grund- , Bestimmungswörter, Fugenelemente (in der Komposition)
- Modifikation von Stämmen (Umlaut, Ablaut)
  - *Mutter, Mütter*
  - *schwimmen, schwamm, geschwommen*
- Morpho-phonologische Prozesse

Morphophonologische Prozesse

- Systematische Modifikation, Einfügung und Tilgung von Lauten/Phonemen, die aus der phonetisch/ phonologischen Struktur vorhersagbar sind:

  Einfügung: *bad+st → badest*
  Tilgung: *ras+st → rast*

- Umlaut und Ablaut sind morphologische Flexionseigenschaften, die für einzelne Wörter im Lexikon spezifiziert werden müssen.
- Morphophonologische Regeln sind Teil der Grammatik (und werden Morphologise-Systemen systematisch behandelt)

Morphologische Verarbeitung in der Computerlinguistik

- Systeme zur morphologischen Analyse werden in vielen computerlinguistischen Anwendungen eingesetzt (z.B. Rechtschreib- und Grammatikkorrektur; Informationszugriff; maschinelle Übersetzung).
- Komponenten:
  - Lemmatisierung: Flexionsmorphologische Analyse: Ermittlung des Stammes/Lemmas („stemming") und ggf. der in den Flexionsformen enthaltenen grammatischen Information
  - Derivativ- und Komposita-Zerlegung: Reduktion komplexer Wörter (Ableitungen und Zusammensetzungen) auf ihre Bestandteile

Ein Flexionsbeispiel aus dem Türkischen

- *Evlerinizdeyiz*
- *Ev+ler+iniz+de+yiz*
- *Haus+pl+poss-2.pers-pl+in+wir-sind*
- *"Wir sind in euren Häusern"*

Ein Kompositionsbeispiel aus dem Deutschen

- Forstspezialrückeschlepper
- Forst+spezial+rücke+schlepper

Bei diesen Fahrzeugen handele es sich nämlich um Forstspezialrückeschlepper, deren Fahren als Beispiel in der Lohngr. W 7 Fallgr. 1 angeführt sei. (...) Ein Forstspezialschlepper, dessen Bestimmung das "Rücken" sei, sei nach den Regeln des allgemeinen Sprachgebrauchs ein Forstspezialrückeschlepper. Dabei spiele es auch keine Rolle, in welcher Reihenfolge die Bestandteile dieses Wortes verwendet seien. Mit Forstspezialrückeschlepper gleichbedeutend wäre auch "Spezialforstrückeschlepper", "Forstrückespezialschlepper" oder "Rückeforstspezialschlepper".

Aus einer Urteilsbegründung des Bundesarbeitsgerichtes

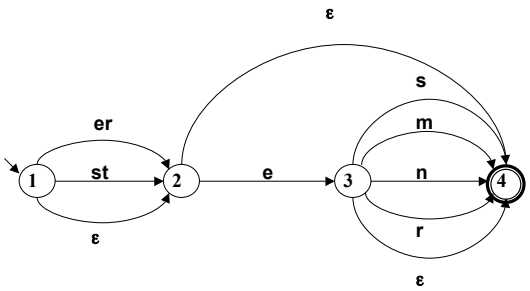
Morphologische Verarbeitung in der Computerlinguistik

- Methodisches Werkzeug für alle Aufgaben morphologischer Verarbeitung sind „Endliche Automaten“.
- Wir betrachten die Verwendung endlicher Automaten an einer vergleichsweise einfachen Teilaufgabe der Lemma-Ermittlung.

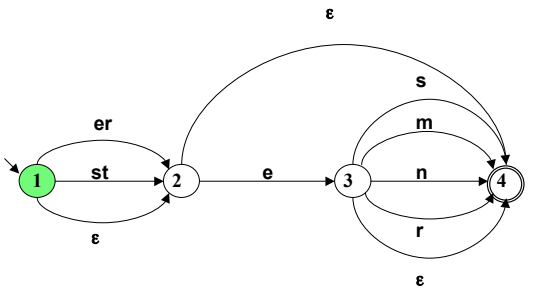
Adjektivflexion: Paradigma
(nur sog. „starke Flexion“)

klein+er	klein+e	klein+es	klein+e
klein+es/en	klein+er	klein+es/en	klein+er
klein+em	klein+er	klein+em	klein+en
klein+en	klein+e	klein+es	klein+e
klein+er+er	klein+er+e	klein+er+es	klein+er+e
klein+er+es/en	klein+er+er	klein+er+es/en	klein+er+er
klein+er+em	klein+er+er	klein+er+em	klein+er+en
klein+er+en	klein+er+e	klein+er+es	klein+er+e
klein+st+ er	klein+st+e	klein+st+ es	klein+st+e
klein+st+es/en	klein+st+er	klein+st+es/en	klein+st+er
klein+st+em	klein+st+er	klein+st+em	klein+st+en
klein+st+en	klein+st+e	klein+st+es	klein+st+e

Adjektivendungen: Darstellung durch Zustandsdiagramm

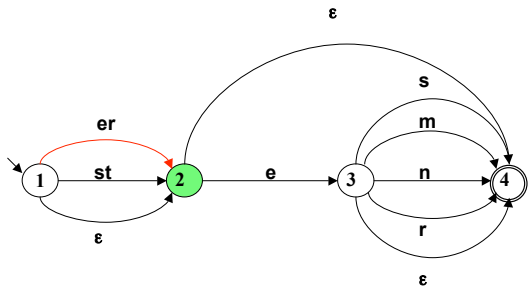


Funktion des Zustandsdiagramms [1]



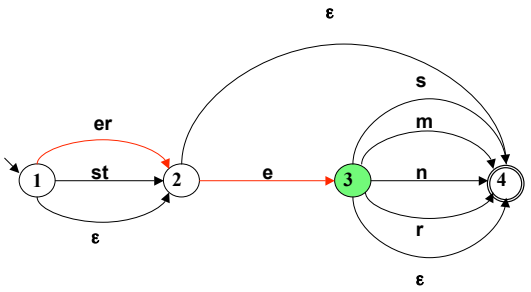
klein eres

Funktion des Zustandsdiagramms [2]



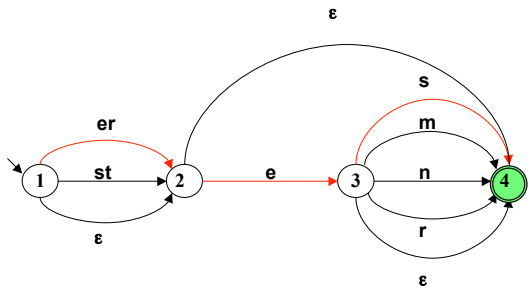
klein er<u>es</u>

Funktion des Zustandsdiagramms [3]



klein er<u>es</u>

Funktion des Zustandsdiagramms [4]



klein eres_

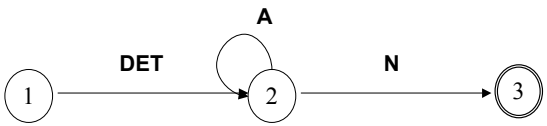
Zustandsdiagramme: Ein zweites Beispiel [1]

- Wortarten kombinieren sich in bestimmter Weise zu Satzteilen
- Um zu testen, ob eine Wortfolge in einem Dokument eine erlaubte Abfolge von Wortarten darstellt, können Zustandsdiagramme benutzt werden.
- Das folgende Zustandsdiagramm akzeptiert bestimmte einfache Nominalausdrücke, wie

  der Wagen

  eine interessante Vorlesung

  das neue schöne rote Dach



Zustandsdiagramme: Ein zweites Beispiel [2]

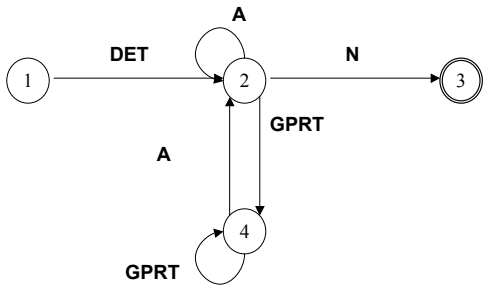
- Das „Alphabet“ des Zustandsdiagramms sind Wortart-Bezeichnungen („Wortart-Tags“ oder auch „POS-Tags“, POS für „part of speech“, engl. für „Wortart“), in unserem Beispiel DET, A, N
  - DET (Artikel)
  - A (adjektivisches Attribut)
  - N (Gattungssubstantiv, Gattungsnomen)
- Erkannte „Worte“ sind erlaubte Abfolgen von Wortartsymbolen, z.B. „DET N“, „DET A N“, „DET A A A N“
- Im Gegensatz zum Adjektivendungsdiagramm akzeptiert das Nominalausdrucksdiagramm beliebig lange Worte und beschreibt eine unendliche Sprache. Grund: Es enthält eine Schleife, es ist zyklisch.

Zustandsdiagramme: Ein zweites Beispiel [3]

Das abgebildete Diagramm akzeptiert auch Adjektive, die mit (einer oder mehreren) Gradpartikeln (GPRT) versehen sind, wie z.B.

eine ziemlich interessante Vorlesung

das recht neue sehr sehr schöne rote Dach



Definitionen: Alphabet und Wort

- Ein Alphabet Σ ist eine endliche, nicht-leere Menge von Symbolen.
- Ein Wort w über dem Alphabet Σ ist eine endliche Kette von Symbolen aus Σ .
- Die Wortlänge |w| eines Wortes w ist die Anzahl der verketteten Symbole von w.
- Das leere Wort ε ist das Wort mit Wortlänge 0 ($|\varepsilon|=0$).

Definitionen: Sprache

- Ein Sprache über dem Alphabet Σ ist eine Menge von Worten über Σ .

  Zwei besondere Sprachen:
- Die leere Wortmenge $\varnothing$ heißt die „leere Sprache“.
- Die maximale Sprache, die die Menge aller Worte über dem Alphabet Σ umfasst, ist Σ^* (der „Stern“ von Σ).

  Anmerkung:

  Für jedes Alphabet Σ gilt: $\varepsilon \in \Sigma^*$.

Beispiele

Beispiel 1:

Σ = {e, m, n, r, s, t}

e, er, rrrrr, mnstmnst, ... $\in \Sigma^*$

L = {ε, e, er, em, en, es, ere, erer, erem, eren, eres, st, ste, stem, sten, ster, stes}

Beispiel 2:

Σ = {DET, A, N}

L = {DET N, DET A N, DET A A N...}

Alternative Formulierung:

L = {DET A^n N| n $\in$ **N**}

Bemerkungen:

- Mit **N** bezeichnen wir hier die Menge der natürlichen Zahlen inklusive 0.
- a^n ist die Kette, die durch n-faches Hintereinander-schreiben des Symbols a entsteht (für n=0 ist $a^n = \varepsilon$)

Beispiele

Beispiel 3:

Σ = {0,1,2,3,4,5,6,7,8,9}

L = {x_1 ... x_n y| n $\in$ **N**, $x_i \in \Sigma$ für $1 \leq i \leq n$, y $\in$ {0,5}, n $\in$ **N**}

(die Menge der durch 5 teilbaren natürlichen Zahlen, wenn wir Ziffernfolgen mit 0-Präfixen ebenfalls zulassen)