Einführung in die Computerlinguistik

Mengen und Formale Sprachen

Laura Kallmeyer
Heinrich-Heine-Universität Düsseldorf

Wintersemester 2011/2012

Überblick

1. Modellbildung
2. Mengen
3. Formale Sprachen

Modellbildung (1)

Um etwas berechnen zu können, benötigt man ein mathematisches Modell davon.

Derartige Modelle

- sind künstlich geschaffen.
- sind materiell oder immateriell.
- sind ein vereinfachtes Abbild.
- sind zweckgerichtet.
- stellen eine Abstraktion dar.
- sind eine Repräsentation.
- beinhalten Modellierungsannahmen.

Modellbildung (2)

Unser Modellierungsgegenstand sind natürliche Sprachen.

Als Modell verwenden wir Formale Sprachen.

- Formale Sprachen sind Mengen von Wörtern (entspricht in natürlichen Sprachen den *Sätzen*).
- Ein Wort ist eine Folge von Zeichen/Symbolen (in natürlichen Sprachen *Wörtern*).
- Was in der Menge ist, ist ein "grammatisch korrektes Wort", alles andere nicht.

Für "strukturierte" formale Sprachen lassen sich endliche Mengen von Regeln/Grammatiken angeben, die diese beschreiben.

Wir gehen davon aus, da alle natürlichen Sprachen durch endlich viele Regeln beschreibbar sind, da wir sie ansonsten nicht sprechen/verstehen könnten.

Mengen (1)

- Eine Menge ist eine Zusammenfassung beliebiger Objekte, genannt Elemente, zu einer Gesamtheit, wobei keines der Objekte die Menge selbst sein darf.
- Zwei Mengen sind gleich, g.d.w. sie die gleichen Elemente enthalten.
- Es gibt genau eine Menge, die keine Elemente enthält, die leere Menge $\emptyset$.

Mengen (2)

Es gibt verschiedene Arten, eine Menge zu beschreiben:

- *Explizite Mengendarstellung* (Auflistung der Elemente):
  $\{a_1, a_2, \ldots, a_n\}$ ist die Menge, die genau die Elemente $a_1, a_2, \ldots, a_n$ enthält.
  Beispiel: $\{2, 3, 4, 5, 6, 7\}$
- *Implizite Mengendarstellung* (Beschreibung der Eigenschaften der Elemente):
  $\{x|A\}$ ist die Menge, die genau die Objekte x enthlt, auf die die Aussage A zutrifft.
  Beispiele: $\{x|x \in \mathbb{N}$ und $x < 8$ und $1 < x\}$,
  $\{x|x$ ist ein deutsches Nomen, das auf *-ung* endet$\}$

Mengen (3)

Beziehungen zwischen Mengen und Elementen:

- Eine Menge N ist eine Teilmenge der Menge M genau dann, wenn alle Elemente von N auch Elemente von M sind.
- Eine Menge N ist eine echte Teilmenge der Menge M genau dann, wenn N eine Teilmenge von M ist und wenn M und N ungleich sind.

Notation

$x \in A$	x ist ein Element von A
$A \subseteq B$	A ist eine Teilmenge von B
$A \subset B$	A ist eine echte Teilmenge von B

Mengen (4)

Operationen auf Mengen:

- Schnitt: $A \cap B = \{x \mid x \in A$ und $x \in B\}$
  Bsp.: $\{1, 2, 3\} \cap \{2, 4, 5\} = \{2\}$
- Vereinigung: $A \cup B = \{x \mid x \in A$ oder $x \in B\}$
  Bsp.: $\{1, 2, 3\} \cup \{2, 4, 5\} = \{1, 2, 3, 4, 5\}$
- Differenz: $A \setminus B = \{x \mid x \in A$ und nicht $x \in B\}$
  Bsp.: $\{1, 2, 3\} \setminus \{2, 4, 5\} = \{1, 3\}$
- Komplement (in U): $C_U(A) = \{x \mid x \in U$ und $x \notin A\}$
  Bsp.: $C_{\{1,2,3,4,5,6\}}(\{2, 4, 5\}) = \{1, 3, 6\}$
  Wenn U feststeht, dann wird das Komplement von A auch mit $\bar{A}$ bezeichnet.

Mengen (5)

Eigenschaften der Mengeoperationen

Kommutativgesetz

$$A \cap B = B \cap A, A \cup B = B \cup A$$

Assoziativgesetz

$$(A \cap B) \cap C = A \cap (B \cap C), (A \cup B) \cup C = A \cup (B \cup C)$$

Distributivgesetz

$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$, $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

de Morgan

$$C_U(A \cap B) = C_U(A) \cup C_U(B), C_U(A \cup B) = C_U(A) \cap C_U(B).$$

Mengen (6)

Die Potenzmenge einer Menge M (Notation $\mathcal{P}(M)$) ist die Menge aller Teilmengen von M, also

$$\mathcal{P}(M) = \{N | N \subseteq M\}$$

Bsp.: $\mathcal{P}(\{a, b\}) = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}$

- Für endliche Mengen gilt: ist M eine n-elementige Menge, so ist $\mathcal{P}(M)$ eine 2^n-elementige Menge.
- Man kann $\mathcal{P}(M)$ dann auch auffassen als die Menge aller möglichen Funktionen, die von M nach $\{0, 1\}$ abbilden.

Formale Sprachen (1)

- Ein Alphabet Σ ist eine nichtleere endliche Menge von Symbolen/Zeichen.

  Bsp.:

  $\Sigma_1 = \{a, b, c, d, e\}$,

  $\Sigma_2 = \{$*der, die, das, Auto, Karl, Maria, schenkt, repariert*$\}$
- Ein Wort über einem Alphabet Σ ist eine endliche Kette/Folge $x_1 \dots x_n$ von Symbolen/Zeichen aus Σ ($n \geq 0$). Das Wort, das aus null Zeichen besteht heißt leeres Wort und wird mit ε bezeichnet.

  Bsp.:

  Wörter über Σ_1: *abbcad*, ε;

  Wörter über Σ_2: *Karl repariert das Auto, Maria schenkt Karl das Auto, Auto das*

Formale Sprachen (2)

- Die Menge aller Wörter über einem Alphabet Σ bezeichnen wir mit Σ^*.
- $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$ ist die Menge der nichtleeren Wörter über Σ.
- Die Länge eines Wortes w (Notation $|w|$) bezeichnet die Gesamtzahl der Zeichen in w.

  Bsp.:

  $|abbaca| = 6$, $|\varepsilon| = 0$,

  $|Karl\ repariert\ das\ Auto| = 4$
- Die Anzahl, mit der ein bestimmtes Zeichen $a \in \Sigma$ in einem Wort $w \in \Sigma^*$ auftritt wird mit $|w|_a$ bezeichnet.

  Bsp.: $|abbaca|_b = 2$

Formale Sprachen (3)

- Die Konkatenation/Verkettung zweier Wrter $u = a_1 a_2 \dots a_n$ und $v = b_1 b_2 \dots b_m$ mit $n, m \geq 0$ ist

$$u \circ v = a_1 \dots a_n b_1 \dots b_m$$

  Häufig schreiben wir uv statt $u \circ v$.

- ε ist das neutrale Element bzgl. Konkatenation: $w \circ \varepsilon = \varepsilon \circ w = w$.
- Assoziativität: $u \circ (v \circ w) = (u \circ v) \circ w$
- w^n: w wird n-mal mit sich selbst verkettet. $w^0 = \varepsilon$: w wird '0-mal' mit sich selbst verkettet.
- Die Umkehrung eines Wortes w wird mit w^R bezeichnet. Bsp.: $(abcd)^R = dcba$.

Formale Sprachen (4)

- Eine formale Sprache L ist eine Menge von Wörtern über einem Alphabet Σ, also $L \subseteq \Sigma^*$.
- Seien $L_1 \subseteq \Sigma^*$ und $L_2 \subseteq \Sigma^*$ zwei Sprachen über dem Alphabet Σ. Dann entstehen durch die Verknüpfung mit Mengenoperatoren neue Sprachen über Σ:

$$L_1 \cup L_2,\ L_1 \cap L_2,\ L_1 \setminus L_2$$

- Die Verkettung von Wörtern kann ausgedehnt werden auf die Verkettung von Sprachen:

$$L_1 \circ L_2 := \{v \circ w \in \Sigma^* | v \in L_1, w \in L_2\}$$

Formale Sprachen (5)

Beispiele formaler Sprachen:

$$\begin{aligned} L_{copy} &= \{ww \,|\, w \in \{a,b\}^*\} \\ &= \{\varepsilon, aa, bb, abab, baba, aaaa, bbbb, \dots\} \end{aligned}$$

$$\begin{aligned} L_{count} &= \{a^n b^n \,|\, n \geq 0\} \\ &= \{\varepsilon, ab, aabb, aaabbb, a^4b^4, \dots\} \end{aligned}$$

$$\begin{aligned} L_{mix} &= \{w \,|\, w \in \{a,b,c\}^*, |w|_a = |w|_b = |w|_c\} \\ &= \{\varepsilon, abc, bac, acb, cab, bca, cba, aabbcc, \dots\} \end{aligned}$$

Formale Sprachen (6)

Wie kann man eine formale Sprache beschreiben?

- Durch eine explizite Angabe der Sprache, z.B. durch Aufzählung der Wörter.
- Durch die Angabe einer Grammatik, die beschreibt, wie man die Wörter der Sprache generieren kann.
- Durch Angabe eines Automaten, der für ein gegebenes Wort überprüft, ob es zur Sprache gehört oder nicht.