Algorithmen zu kontextfreien Grammatiken

Masterseminar SoSe 2015
Algorithmische und formale Aspekte II b

Martin Hofmann, Institut für Informatik
Hans Leiß, Centrum für Informations- und Sprachverarbeitung
Universität München

SoSe 2015: Mi, 12-14 Uhr, Raum C003, Oettingenstr.67

(ggf. auch im Bachelorstudium)

13. April 2015

Zeitplan

Woche	Termin	Wer	Thema
1	15.4.	L/H	Themenvergabe
2-H	22.4.		CYK-Erkenner
3	29.4.		Earley-Erkenner
4	6.5.		Valiant: CFG-Erkennen durch MM
5	13.5.		Lee: MM durch CFG-Erkennen
6	20.5.		Erkenner für datenabh.Grammatiken
7	27.5.		Axiomatisierung mit Fixpunkten
8	3.6.		Axiomatisierung mit Division
9	10.6.		Programmanalyse und Sprachfaktoren
10	17.6.		2.Stufige Axiomatisierung von Sprachfamilien
11	24.6.		Unlernbarkeitssätze von Gold und Kracht
12-H	1.7.		Verband der Syntaktischen Konzepte
13	8.7.		Limesidentifizierung gewisser CFLs
14-L	15.7.		Lernen mit Semantik

I. Parsen mit kontextfreien Grammatiken

1. CYK-Erkenner und Binärform (mit Effizienzfragen)
   - Sippu/Soisalen-Soininen: Parsing Theory I. EATCS Monographs in CS (1988)
   - M.Lange, H.L.: To CNF or not to CNF? Informatica Didactica 8, 2009 (online)

   Erkennungsproblem: $\{(w, G) \in \Sigma^* \times CNF \mid w \in L(G)\}$
   Normalformberechnung: $G \in CFG \mapsto G' \in CNF \mid 2NF$
   Löst das Erkennungsproblem für CNF-CFGs in $O(|G||w|^3)$
   Berechne $G \mapsto G'$ in $O(|G|^2)$ für CNF bzw. $O(|G|)$ für 2NF

2. Earley-Erkenner zum Parsen mit kontextfreien Grammatiken
   - Lehrbücher (Hopcroft/Ullman, Rich, Naumann/Langer)
     M.Harrison: Intr.to Formal Languages. Addison Wesley, 1978
   - Originalarbeit J.Earley (1967)
   - Graham/Harrison/Ruzzo: An improved context-free recognizer.
     ACM Transactions on Programming languages (1980)

   Erkennungsproblem: $\{(w, G) \in \Sigma^* \times CFG \mid w \in L(G)\}$
   (Implizite Normalform durch gespaltene Regeln $A \to \alpha \bullet \beta$)
   Berechnung von Hilfsrelationen, $\{(A, B) \in N \times N \mid A \Rightarrow^* B\}$

   Löst das Erkennungsproblem für CFGs in $O(|G|^2|w|^3)$

3. Valiant-Erkenner und Multiplikation Boole'scher Matrizen
   - J.-P.Bernardy, K.Claessen: Efficient Divide-and-Conquer Parsing of Practical Context-Free Langugages (2013) (`PP.pdf`)
   - R.Bird, O.de Moor: Algebra of Programming (sec. 4.2) Prentice-Hall, 1997

   Erkennungsproblem: $\{(w, G) \in \Sigma^* \times CNF \mid w \in L(G)\}$
   Reduktion des Erkennens von w auf die Berechnung der Matrix

$$M(w)_{i,j} = \{A \in N \mid A \Rightarrow^* w[i..j]\}$$

   als transitive Hülle $M = I(w)^+$ der Anfangsmatrix

$$I(w)_{i,j} = \begin{cases} \{A \in N \mid A \Rightarrow w[i..j]\}, & j = i+1 \\ \emptyset, & j \neq i+1 \end{cases}$$

   Löst das Erkennen für eine CNF-Subklasse in $O(|G| \cdot \log^3 |w|)$,
   Valiant für CNF-CFGs in $O(|G| \cdot |w|^{2.8}) < O(|G| \cdot |w|^3)$

4. Reduktion der Matrix-Multiplikation auf CFG-Erkennung
   - L.Lee: Fast Context-Free Grammar Parsing Requires Fast Boolean Matrix Multiplication. Journal of the ACM, vol.49, no.1, 1–15 (2002)

   Reduktion der Multiplikation von 0-1-Matrizen auf das Erkennen mit kontextfreien Grammatiken. (`p1-lee-1.pdf`)

   Valiant $\oplus$ Lee ergeben die algorithmische Äquivalenz der Probleme
   - *BMM* := Multiplikation Boole'scher Matrizen und
   - Erkennung mit kontextfreien Grammatiken

5. Ein Parsergenerator für nicht-kontextfreie Sprachen wie XML/HTML

   - T.Jim, Y.Mandelbaum, D.Walker: Semantics and Algorithms for Data-dependent Grammars. Proceedings POPL 2010

   Parsergeneratoren für Sprachen zur Web-Entwicklung, Systemprogrammierung, Netzwerkprogrammierung, die z.B. XML-Tags `<label> ... <\label>`, Längenangaben für Binärstrings, Variablenbelegung u.ä. nicht-kontextfreie Mittel verwenden.
   Verwendet Transduktoren und Erweitert einen Earley-Parser

II. Axiomatisierungsfragen und Fixpunkte

6. Axiomatisierung der Gleichungstheorie kontextfreier Sprachen:
   - N.B.Grathwohl, F.Henglein, D.Kozen: Infinitary Axiomatization of the Equational Theory of Context-Free Languages. Proc.FICS 2013 (`fics2013.univ-mlv.fr/papers.html`)

   Eine kontextfreie Grammatik ist ein Ungleichungssystem, dessen kleinste Lösung die definierten Sprachen sind, z.B. hat

   $$X \geq aYYb, \quad Y \geq aYb + c$$

   die kleinste Lösung: $X = aYYb, Y = \{a^n cb^n \mid n \in \mathbb{N}\}$.

   Eine *Chomsky-Algebra* ist ein idempot.Halbring $(A, +, \cdot, 0, 1)$, wo jedes solche System $\vec{x} \geq \vec{p}(\vec{x})$ eine kleinste Lösung hat. Die Lösung kann man mit *Fixpunktausdrücken* $\mu x.p(x)$ darstellen.

   Welche Axiome reichen, um alle (Fixpunkt-) Gleichungen $r = s$ zu beweisen, die im Bereich $\mathcal{C}\Sigma^*$ aller kontextfreien Sprachen wahr sind? Memo: Das Problem ist unentscheidbar:

   $$\{(G_1, G_2) \in CFG \times CFG \mid L(G_1) = L(G_2)\}$$

7. Andere Axiomatisierung: Reguläre Algebren $(A, +, \cdot, \backslash, /, 0, 1)$ mit Divisionen

   - R.Backhouse: Regular algebra applied to language problems. Journal of Logic and Algebraic Programming, 2004

Eine *Galois-Verbindung* $g : \mathcal{A} \rightleftarrows \mathcal{B} : f$ zwischen partiellen Ordnungen $\mathcal{A} = (A, \leq^A)$ and $\mathcal{B} = (B, \leq^B)$ ist ein Paar (g, f) von Funktionen, so daß für $a \in A, b \in B$,

$$f(b) \leq^A a \iff b \leq^B g(a). \tag{1}$$

Backhouse nennt $(A, +, \cdot, \leq, 0, 1)$ regulär, wenn $(A, +, 0, \leq)$ vollständig ist und für alle $c \in A$ die $\lambda y(c \cdot y)$ und $\lambda x(x \cdot c)$ der f-Teil einer Galois-Verbindung von $(A, \leq)$ mit sich sind, d.h. entsprechende Divisionen existieren, z.B.

$$\lambda x(c \backslash x) = g_c : \mathcal{A} \rightleftarrows \mathcal{B} : f_c = \lambda y(c \cdot y)$$

$$f_c(b) = c \cdot b \leq^A a \iff b \leq^B c \backslash a = g_c(a)$$

Backhouse gibt eine Reihe von Konstruktionen an, wie man solche Algebren mit Divisionen aus vorhandenen (z.B. den Regulären Sprachen) aufbaut und in der Informatik anwendet, z.B. auf

- Berechnung kürzester Wege in Graphen
- Berechnung von Eigenschaften kontextfreier Grammatiken
- CYK-Parsing, Editierabstand
- Subsumption $L(G) \subseteq L$ für kontextfreie G, beliebige L als Fixpunktbedingung über alle Faktoren $M \backslash L / N$ von $L(G)$

Comp.Ling.: Sprachdivisionen L/M, $M \backslash L$: vgl. Kategorialgrammatik

Memo: Die regulären Sprachen bilden eine solche residuierte Algebra, die kontextfreien aber nicht: für kontextfreie L und Wörter u, v ist $L/\{u\}$ kontextfrei, aber $L/\{u, v\}$ i.a. nicht.

(Nachteil: Backhouse benutzt eine bescheuerte Notation)

8. Anwendung der Faktortheorie bei regulären Sprachen

   - O. de Moor, St. Drape, D. Lacey, G. Sittampalam: Incremental Program Analysis via Language Factors. (2004) `http://web.comlab.ox.ac.uk/oucl/work/oege.de.moor/opubs.html`

   Für kontextfreie G und reguläre Schranke R kann man das Obere-Schranke-Problem

   $$\{(G, R) \in CFG \times Reg \mid L(G) \subseteq R\}$$

   effektiv lösen. Man berechnet eine Matrix aller Sprachfaktoren $N \backslash L(G) / M$ und kann durch Induktion über die Matrixdimension Eigenschaften von Programmen prüfen.

   (Die Arbeit war viel lesbarer als Backhouse's Verallgemeinerung, finde sie aber im Moment nicht als .pdf.)

III. Grammatiklernen

9. Grammatiklernen aus positiven Daten: Unlernbarkeitssätze

   - E.Gold: Language identification in the limit. Information and Control 10, 447-474 (1967)
   - D.Angluin: Inductive Inference of formal languages from positive data. Information and Control, 21, 46-62 (1980)
   - A.Kornai: Mathematical Linguistics, Kap. 7.3. Springer 2008

„$\mathcal{L}$ ist iitl“: Aus einer Aufzählung $L = \{w_n \mid n \in \mathbb{N}\}$ einer Sprache L der Klasse $\mathcal{L}$ soll eine Folge von Grammatiken G_n (eines zu $\mathcal{L}$ passenden Formats) konstruiert werden, die irgendwann konstant wird bei einer Grammatik für L.
Gold zeigt, daß das für jedes $\mathcal{L} \supset$ endliche Sprachen unmöglich ist, Kracht (s.Kornai) zeigt es (via ein Kriterium von Angluin) für $\mathcal{L} =$ reguläre nicht-zählende Sprachen, i.e. solche mit

$$vx^4w \in L \iff vx^5w \in L.$$

10. Lernen kontextfreier Grammatiken, Syntaktische Konzepte

- A.Clark: A learnable representation for syntax using residuated lattices. Proc. FG 2009, Springer LNCS 5591, 2011, pp 183-198

Jedes $L \subseteq \Sigma^*$ liefert eine eine Galois-Verbindung

$$^{\triangleright} : (\Sigma^*, \subseteq) \rightleftarrows (\Sigma^* \times \Sigma^*, \supseteq) : {}^{\triangleleft}$$

zwischen Wort- und Kontextmengen, $A \subseteq \Sigma^*$, $C \subseteq \Sigma^* \times \Sigma^*$,

$$\begin{aligned} A^{\triangleright} &:= \{(u,v) \in \Sigma^* \times \Sigma^* \mid uAv \subseteq L\} \\ C^{\triangleleft} &:= \{w \in \Sigma^* \mid \forall (u,v) \in C : uwv \in L\}, \end{aligned}$$

wobei $A \subseteq A^{\triangleright\triangleleft}$ und $C \subseteq C^{\triangleleft\triangleright}$. *Syntaktische Konzepte* von L sind die (A, C) mit $A^{\triangleright} = C$, $C^{\triangleleft} = A$, also $A = A^{\triangleright\triangleleft}$, $C = C^{\triangleleft\triangleright}$. Die Konzepte von L bilden einen Verband V,

$$V \simeq \mathcal{P}(\Sigma^*)/\equiv_L \text{ für } X \equiv_L Y :\iff X^{\triangleright} = Y^{\triangleright},$$

der für gewisse kontextfreie L zum Lernen einer CFG für L ausreicht, wenn man eine Aufzählung und ein Orakel für L hat.

11. Lernen von Bedeutungen

- D.Angluin, L.Becerra-Bonache: Learning Meaning before Syntax. ICGI 2008, Springer LNCS 5278, 1-14

Es wird ein Lernmodell vorgeschlagen, bei dem aus mehreren (Situation,Äußerung)-Paaren die Bedeutung von Wörtern der Äußerungen als Objekte oder Relationen zwischen Objekten der Situationen gelernt werden.
Konstruiert einen endlichen Transduktor, der Wörter auf Begriffe abbildet; soll das Lernen von 2-Wort-Sätzen bei Kleinkindern verständlich machen.