

アルゴリズムとデータ構造III 3回目:10月21日

構文解析 CYK法

授業資料 <http://ir.cs.yamanashi.ac.jp/~ysuzuki/public/algorithm3/index.html>

授業の予定(中間試験まで)

1	10/07	スタック(後置記法で書かれた式の計算)
2	10/14	チューリング機械, 文脈自由文法
3	10/21	構文解析 CYK法
4	11/04	構文解析 CYK法
5	11/04	構文解析(チャート法), グラフ(ダイクストラ法)
6	11/11	構文解析(チャート法), グラフ(ダイクストラ法, DPマッチング)
7	11/18	グラフ(DPマッチング, A*アルゴリズム)
8	11/25	グラフ(A*アルゴリズム), 前半のまとめ
9	12/02	中間試験

10/28の代わりに補講日は後日相談

授業の予定(中間試験以降)

10	12/09	全文検索アルゴリズム(simple search, KMP)
11	12/16	全文検索アルゴリズム(BM, Aho-Corasick)
12	01/06	全文検索アルゴリズム(Aho-Corasick), データ圧縮
13	01/13	暗号(黄金虫, 踊る人形) 符号化(モールス信号, Zipfの法則, ハフマン符号)テキスト圧縮
14	01/20	テキスト圧縮(zip), 音声圧縮(ADPCM, MP3, CELP), 画像圧縮(JPEG)
15	02/03	期末試験

本日のメニュー

- 構文解析(先週の復習)
 - 構文木(急いで走る一郎を見た)
 - 代表的な構文解析アルゴリズム
- CYK法
 - CYKアルゴリズムの説明
 - 解析例(急いで走る一郎を見た)
 - Step by step で解析

構文解析 CYK法

- 文脈自由文法で生成された文から自動的に構文木を生成する。

構文解析とは(Wikipediaより)

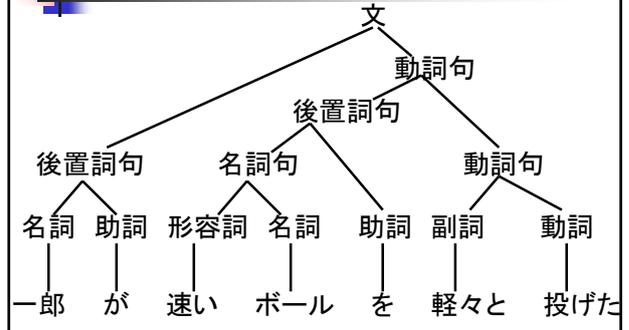
- ある文章の文法的な関係を説明すること(parse)。計算機科学の世界では、構文解析は字句解析(Lexical Analysis)とともに、おもにプログラミング言語などの形式言語の解析に使用される。また、自然言語処理に応用されることもある。
- コンパイラにおいて構文解析を行う機構を構文解析器(Parser)と呼ぶ。
- 構文解析は入力テキストを通常、木構造のデータ構造に変換し、その後の処理に適した形にする。字句解析によって入力文字列から字句を取り出し、それらを構文解析器の入力として、構文木や抽象構文木のようなデータ構造を生成する。

構文解析アルゴリズム

- 入力文(記号列)が与えられたとき, 文法によってその文を解析し, その構造を明らかにする
- 代表的な構文解析アルゴリズム
 - ▶ **CYK法**
 - チャート法
 - アーリー法
 - LR法

構文木の例

(一郎が速いボールを軽々と投げた)



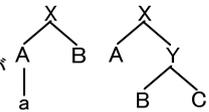
CYK (Cocke-Younger-Kasami) 法

- ボトムアップアルゴリズム
- 扱える文法
 - チョムスキーの標準形
 - $A \rightarrow BC$ (2分木)
 - $A \rightarrow a$
 - $A \rightarrow BCD$, $A \rightarrow aB$ などは直接は扱えない
- CYK表
 - 構文解析の途中経過を保持するための表

CYKアルゴリズム

- チョムスキーの標準形で表される文脈自由文法を対象とした構文解析法
- チョムスキーの標準形
 - $A \rightarrow BC$ ($A, B, C \in V_n$)
 - $A \rightarrow a$ ($A \in V_n, a \in V_t$)

$X \rightarrow aB$ はチョムスキーの標準形ではないが
 $X \rightarrow AB, A \rightarrow a$ に分解できる
 $X \rightarrow ABC$ はチョムスキーの標準形ではないが
 $X \rightarrow AY, Y \rightarrow BC$ に分解できる



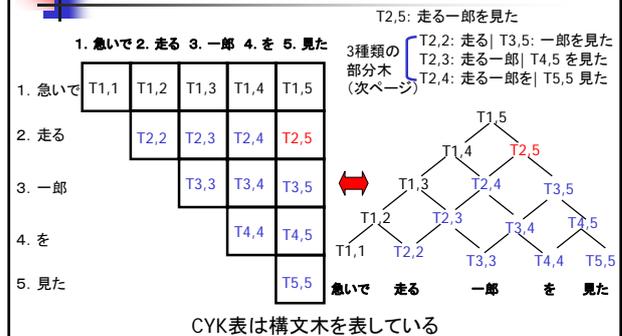
チョムスキーの標準形の例 「急いで走る一郎を見る」

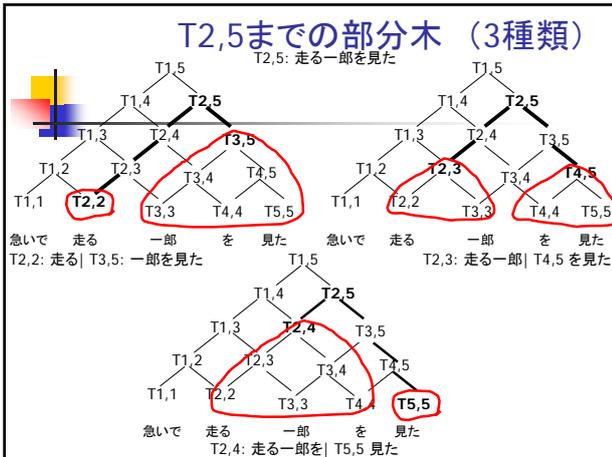
$A \rightarrow BC$ 型

$A \rightarrow a$ 型

- (1) $s \rightarrow pp\ v$
- (2) $s \rightarrow adv\ vp$
- (3) $vp \rightarrow pp\ v$
- (4) $vp \rightarrow adv\ v$
- (5) $np \rightarrow vp\ n$
- (6) $np \rightarrow v\ n$
- (7) $pp \rightarrow np\ p$
- (8) $pp \rightarrow n\ p$
- (9) $adv \rightarrow 急いで$
- (10) $n \rightarrow 一郎$
- (11) $p \rightarrow を$
- (12) $v \rightarrow 走る$
- (13) $v \rightarrow 見る$

CYK構文解析の概要





CYKアルゴリズム

- $A \rightarrow a$ の生成規則を用いて、主対角線上の要素を計算
for $i=1$ to N
 $T_{i,i} = \{A \mid A \rightarrow w_i\}$
- $A \rightarrow BC$ の生成規則を用いて、2番目以降の対角線上の要素を計算
for $n=1$ to $N-1$
for $i=1$ to $N-n$
$$T_{i,i+n} = \bigcup_{j=1}^n \{A \mid A \rightarrow BC, B \in T_{i,j}, C \in T_{i+j,i+n}\}$$
- $S \in T_{1,N}$ であれば、 $w_1 \dots w_N$ は開始記号 S から導出可能

CYK構文解析表(完成)

	1. 急いで	2. 走る	3. 一郎	4. を	5. 見た
1. 急いで	adv→急いで	vp→adv v	np→vp n	pp→np p	vp→pp v s→pp v s→adv vp
2. 走る		v→走る	np→v n	pp→np p	vp→pp v s→pp v
3. 一郎			n→一郎	pp→n p	vp→pp v s→pp v
4. を				p→を	
5. 見た					v→見た

- (1) s→pp v
- (2) s→adv vp
- (3) vp→pp v
- (4) vp→adv v
- (5) np→vp n
- (6) np→v n
- (7) pp→np p
- (8) pp→n p
- (9) adv→急いで
- (10) n→一郎
- (11) p→を
- (12) v→走る
- (13) v→見る

CYK構文解析表(1/5) A→a型

	1. 急いで	2. 走る	3. 一郎	4. を	5. 見た
1. 急いで	adv→急いで				
2. 走る		v→走る			
3. 一郎			n→一郎		
4. を				p→を	
5. 見た					v→見た

- (1) s→pp v
- (2) s→adv vp
- (3) vp→pp v
- (4) vp→adv v
- (5) np→vp n
- (6) np→v n
- (7) pp→np p
- (8) pp→n p
- (9) adv→急いで
- (10) n→一郎
- (11) p→を
- (12) v→走る
- (13) v→見る

CYK構文解析表(2/5)

	1. 急いで	2. 走る	3. 一郎	4. を	5. 見た
1. 急いで	adv→急いで	vp→adv v			
2. 走る		v→走る	np→v n		
3. 一郎			n→一郎	pp→n p	
4. を				p→を	
5. 見た					v→見た

- (1) s→pp v
- (2) s→adv vp
- (3) vp→pp v
- (4) vp→adv v
- (5) np→vp n
- (6) np→v n
- (7) pp→np p
- (8) pp→n p
- (9) adv→急いで
- (10) n→一郎
- (11) p→を
- (12) v→走る
- (13) v→見る

CYK構文解析表(3/5)

	1. 急いで	2. 走る	3. 一郎	4. を	5. 見た
1. 急いで	adv→急いで	vp→adv v	np→vp n		
2. 走る		v→走る	np→v n	pp→np p	
3. 一郎			n→一郎	pp→n p	vp→pp v s→pp v
4. を				p→を	
5. 見た					v→見た

- (1) s→pp v
- (2) s→adv vp
- (3) vp→pp v
- (4) vp→adv v
- (5) np→vp n
- (6) np→v n
- (7) pp→np p
- (8) pp→n p
- (9) adv→急いで
- (10) n→一郎
- (11) p→を
- (12) v→走る
- (13) v→見る

CYK構文解析表(4/5)

	1. 急いで	2. 走る	3. 一郎	4. を	5. 見た
1. 急いで	adv→急いで	vp→adv v	np→vp n	pp→np p	
2. 走る		v→走る	np→v n	pp→np p	vp→pp v s→pp v
3. 一郎			n→一郎	pp→n p	vp→pp v s→pp v
4. を				p→を	
5. 見た					v→見た

A→BC型
 (1) s→pp v
 (2) s→adv vp
 (3) vp→pp v
 (4) vp→adv v
 (5) np→vp n
 (6) np→v n
 (7) pp→np p
 (8) pp→n p
 (9) adv→急いで
 (10) n→一郎
 (11) p→を
 (12) v→走る
 (13) v→見る

CYK構文解析表(5/5)

	1. 急いで	2. 走る	3. 一郎	4. を	5. 見た
1. 急いで	adv→急いで	vp→adv v	np→vp n	pp→np p	vp→pp v s→adv vp
2. 走る		v→走る	np→v n	pp→np p	vp→pp v s→pp v
3. 一郎			n→一郎	pp→n p	vp→pp v s→pp v
4. を				p→を	
5. 見た					v→見た

A→BC型
 (1) s→pp v
 (2) s→adv vp
 (3) vp→pp v
 (4) vp→adv v
 (5) np→vp n
 (6) np→v n
 (7) pp→np p
 (8) pp→n p
 (9) adv→急いで
 (10) n→一郎
 (11) p→を
 (12) v→走る
 (13) v→見る

CYK構文解析表(完成！)

	1. 急いで	2. 走る	3. 一郎	4. を	5. 見た
1. 急いで	adv→急いで	vp→adv v	np→vp n	pp→np p	vp→pp v s→pp v s→adv vp
2. 走る		v→走る	np→v n	pp→np p	vp→pp v s→pp v
3. 一郎			n→一郎	pp→n p	vp→pp v s→pp v
4. を				p→を	
5. 見た					v→見た

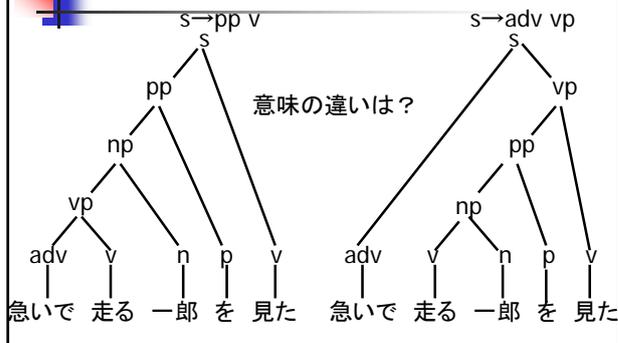
CYK構文解析表 → 構文木 (s→pp v の構文木)

	1. 急いで	2. 走る	3. 一郎	4. を	5. 見た
1. 急いで	adv→急いで	vp→adv v	np→vp n	pp→np p	s→pp v
2. 走る		v→走る	np→v n	pp→np p	
3. 一郎			n→一郎	pp→n p	
4. を				p→を	
5. 見た					v→見た

CYK構文解析表 → 構文木 (s→adv vp の構文木)

	1. 急いで	2. 走る	3. 一郎	4. を	5. 見た
1. 急いで	adv→急いで				s→adv vp
2. 走る		v→走る	np→vp n	pp→np p	vp→pp v
3. 一郎			n→一郎	pp→n p	
4. を				p→を	
5. 見た					v→見た

文脈自由文法に基づく構文木(2種類)



質問: CYK法ではなぜチョムスキー
の標準形しか扱えない?

本日のまとめ

- 構文解析(先週の復習)
 - 構文木(急いで走る一郎を見た)
 - 代表的な構文解析アルゴリズム
- CYK法
 - CYKアルゴリズムの説明
 - 解析例(急いで走る一郎を見た)
 - Step by step で解析