言語解析理論

講師　竹内孔一

本日の内容

- 文の繰り返し構造を扱う理論
- 形式的な言語の扱い方
  - 文脈自由(型)文法
- 文法クラスの位置付け
  - チョムスキー階層

生成できる言語の違い

- もう一つの数学モデル
  - オートマトン(automaton)

言語研究の相関

語・語彙に関する研究

固有表現抽出
専門用語抽出

複合語解析

多言語辞書

語彙意味

音声認識

一文に対する解析

形態素解析

構文解析

意味解析

翻訳

知識処理

文書の解析

要約

質問応答システム

情報検索

topic tracking

表層 ⟷ 深層

時系列から知識の抽出

言語の理論

- 形式言語
  - 形式的な言語の側面を捉える
  - 言葉の意味は捨象して形式に注目した語の性質

  →入れ子の関係（句）
- 句構造文法
  - 例) 私は　その本を　読んだ

    彼の姉は　この本を　読んでました

    刑事は　彼がその犯人であることを　知った

    (主語)(助詞)（目的語)(助詞)（述語)

文脈自由文法

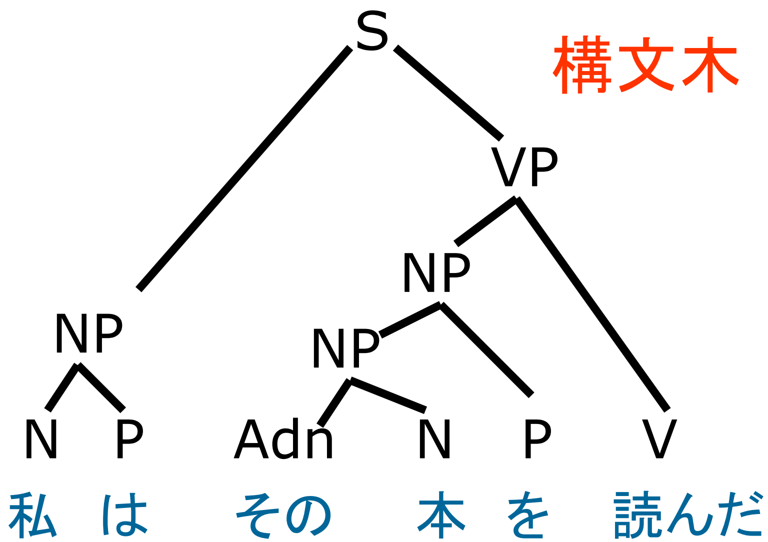
- 形式化
  - 入れ子構造を捉える

非終端記号の集合 V_N

終端記号の集合 V_T

生成規則の集合 P

初期記号 σ



$$G = \langle V_N, V_T, P, \sigma \rangle$$

$$V_N = \{S, NP, VP, P, Adn, V\}$$

V_T = {私, は, その, 本, を, 読んだ}

$$\sigma = S$$

P={S→NP VP, NP→ N P,
NP → NP P, NP → Adn N,
VP → NP V,
N → 私, N→ 本, ..., P→を}

練習3

- 次の文の構文木を完成させ生成規則を記述せよ

  例文)「彼はフランスに行った」

  – ただしVt={彼, は, フランス, に, 行った}とする

  – Vn, σ, P は各自決めること

文脈自由文法

- 理論的な特徴と限界
  - 文法の記述に制限がある

    P について　$A \to \beta \qquad A \in V_N \ \beta \in (V_N \cup V_T)^+$

    - 非終端記号がどこでも書き換え可能→ 文脈自由

    (参考) NP → N P などすべてこの範囲

$\boxed{\text{生成規則の形を制限 → 生成する言語が異なる}}$

4種類の規則に分類　Chomsky 1959
→ 各文法の限界を示した
　チョムスキー階層

チョムスキー階層

- 生成規則による言語の能力の違い

能力大

文法　<=等価=>　オートマトン

0型言語　0型文法　チューリングマシン（汎用計算機）

$$\alpha \to \beta$$

1型言語　文脈依存文法　線形拘束オートマトン

$$\alpha \to \beta \quad |\alpha| \leq |\beta|$$

2型言語　文脈自由文法　プッシュダウンオートマトン

$$A \to \beta \quad A \in V_N \; \beta \in V^+$$

3型言語　正規文法　有限オートマトン

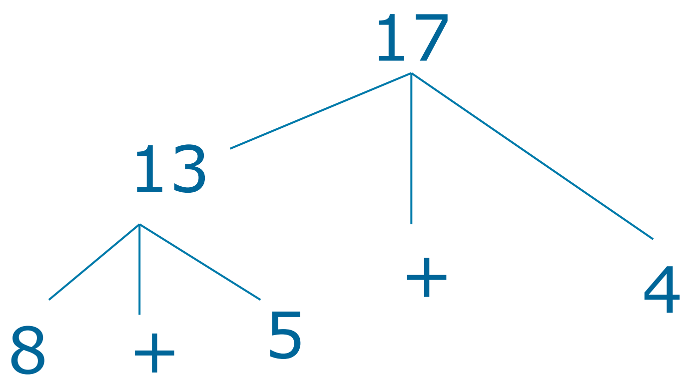
$$A \to aB \qquad A, B \in V_N$$

$$A \to a \qquad a \in V_T^+$$

能力小

問題

- 8+5+4 といった和の計算のための文法はチョムスキー階層のどれにあたるか?
- 文脈自由文法
  - Vt = {8,5,4,+}
  - Vn={ 13 -> 8 + 5, 17 -> 13 + 4}



計算の本質は書き換え

能力の違い

- 以下の2つを取り上げて説明

文法と生成言語について説明する

- 正規文法 regular grammar
  - 有限オートマトン
    finite automaton (fa)

- 文脈自由文法 context free grammar
  - プッシュダウンオートマトン
    push down automaton (pda)

正規文法

- 最も制約された文法 (regular grammar)
  - 生成規則

$$A \rightarrow aB \quad A \rightarrow a \qquad A, B \in V_N \quad \alpha \in V_T^+$$

    - ポイント a (文字) を必ず含む
  - 規則と生成する文字列との関係

例)「平成□□年」を受け付ける正規文法

V_N ={S, Y, A}

V_T ={平成, 年, 1,2,.., 27}

σ= S

単語ベース

P={S→平成 Y,
Y → 1 A,
Y → 2 A,
⋮
Y→ 27 A,
A → 年}

練習4

- 正規文法
  - 「平成□□年」を受け付ける正規文法を記述せよ

    ただし文字ベースとする

    V_T={平, 成, 年, 0,1,2,.., 9}

正規表現との対応は?

- 正規表現を正規文法で表すと

例） $a^{*} = \{\varepsilon, a, aa, aaa, aaaa, \ldots\}$

この場合正規文法は A->ε, A->aA

下記の場合は??

$$a^{+} = \{a, aa, aaa, aaaa, \ldots\}$$

$$(a+b)^{+} = \{a, b, aa, aaa, bb, bbb, ab, abb, \ldots\}$$

過剰生成について

- 受理する目的で作成
  - それ以外のものも受け付けることが多い

例）「平成□□年」を受理する正規文法で記述する

$$G = \langle V_N, V_T, P, \sigma \rangle$$

V_T ={平, 成, 年, 0,1,2,.., 9}

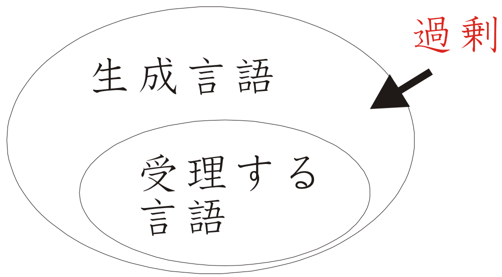
V_N ={S}

σ= S

P={S → 平 S, S → 成 S,
S → 0 S, S → 1 S,
S → 2 S, ..., S → 9 S,
S → 年}

文法Gが生成(受理)する
言語L(G)の一例として

平成05年, 平成119年. . ← 過剰生成



有限オートマトン

- 特徴
  - 内部の状態と外部からの入力で次の状態を決める機械
- 種類
  - 決定性有限オートマトン
    (deterministic finite automaton (fa, dfa))
    有限オートマトンというと普通はこちら
  - 非決定性有限オートマトン
    (non-deterministic finite automaton (nfa))
    → nfa は fa に書き下すことができる
- 重要な関係
  - 正規文法と等価な言語生成能力

有限オートマトン

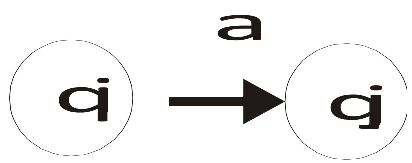
- 定義

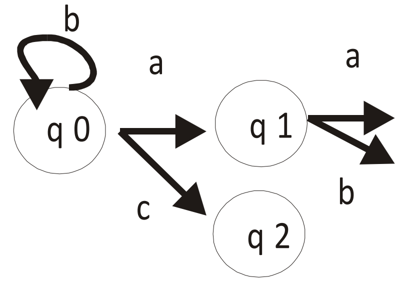
  状態集合をK={q0,q1,q2.., qn}とすると

  - 状態の集合 *K*
  - 入力文字の集合 *S*
  - 状態遷移関数の集合 *P*
  - 初期状態 *q0*
  - 最終状態の集合 *F*

  M=<K,S,P,q0,F>

$$\delta(q_i, a) = q_j$$



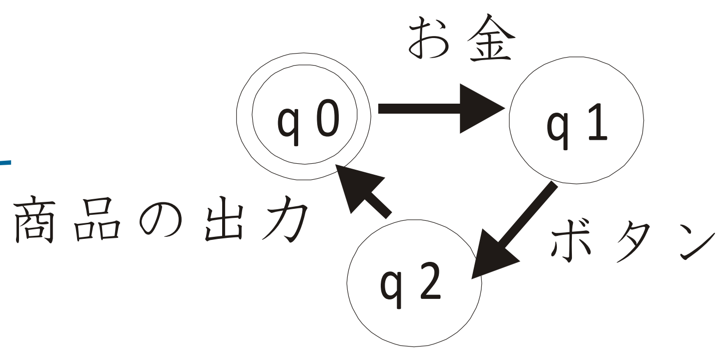


文字a を受け付けて状態qiからqjに遷移する
→ある手順での決まった規則を書くのに適している

- 応用例
  - 自動販売機のモデル

    お金を入れる→ボタンを押す

    → 商品がでる



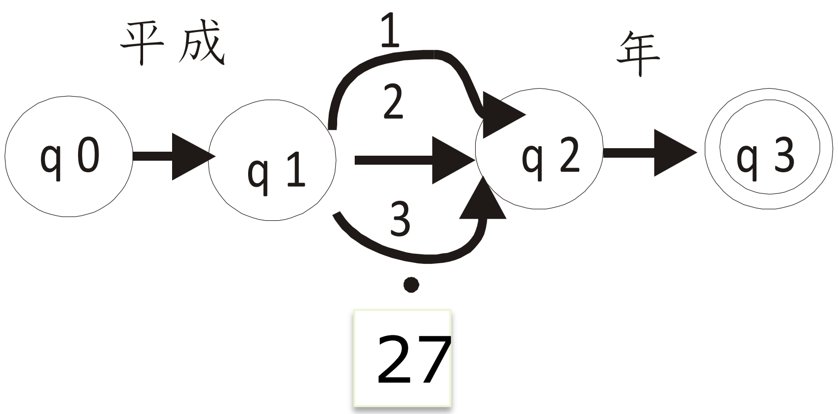
有限オートマトン

- 例題
  - 「平成□ □年」を受理するオートマトンを作成せよ

  解答例

  - 単語ベースの場合

    S = {平成, 1,2,3,..,27, 年}



正規文法

P={S→平成　Y,
Y → 1　A,
Y → 2　A,

Y→ 27　A,
A → 年}

最終状態q3に行かなければ不受理

練習5

- 有限オートマトン
  - 文字ベースS={平, 成, 0,1,2,..,9,年}として「平成□□年」を受理するオートマトンを作成せよ
    (過剰生成しないモデルを目指して)