言語解析論

講師　竹内孔一

この講義の枠組み

- 講義の目標
  - 人間が使う言語をコンピュータ上で扱う
    - 難しさ (人間の情報処理能力の高さ)を知る
    - アプローチを知る
    - モデルを理解する
  - 得られる効果
    - 言語解析の方法，処理モデルの理解
    - 現実的に解析できるツールの使いこなし
  - 評価
    - 試験(5)とレポート(4)と練習問題(1)

講義の資料

- わかりやすい参考書
  - 自然言語処理の基礎　奥村学 コロナ社
- 参考資料
  - 単語と辞書 岩波書店
  - 自然言語処理 IT text (ohmsha)
  - 言語処理のための機械学習入門 高村 大也
- 配布資料
  - 下記のスライド
- スライド
  - http://www.cl.cs.okayama-u.ac.jp/
  - pdf 形式で資料を置く

今日の内容

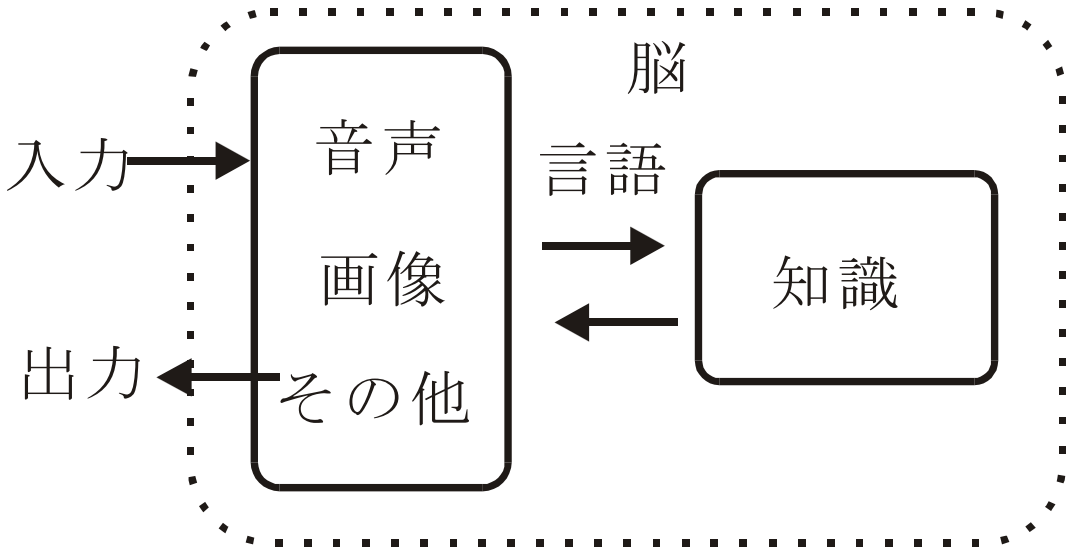
- 言語処理とは
  - 人間が行う言語処理のむずかしさ
    - 脳の処理としての位置付け
  - コンピュータが行う言語処理
    - 計算と知識そのものが言語に関係
- 学問的な背景
  - 文学部言語・哲学科(言語学)
  - 工学部情報工学科(自然言語処理)

  分野を超えた考え方
- 形式言語

脳の言語処理

- 言語のすばらしさ
  - 柔軟な知識変換能力
    - (例) 絵や状況を言葉に置き換える能力

→ 外界と知識とのインターフェース



コンピュータでの言語処理

- 2つの言語処理
  - 人間の言葉 → 自然言語
  - コンピュータ → 形式言語
    - C言語, Java, Perl, HTML
  - 人間の言語処理モデルを参考に作成
    - もとは同じ理論（N. Chomsky）

すべて言語で動いている

- 言語処理を研究する必要性
  - 知的な処理をするために言語が必要

質問

- 共通することは?
  - 秘書に対して何か作業をお願いする
  - コンピュータにC言語でプログラムを書き作業を実行させる

言語処理に対する要求

- 現状
  - 全世界規模で internet による接続
  - 有益な情報が多く存在する
  - どうさがすか?
- 必要性
  - 必要な情報を正確に取り出したい
    - google は keyword base 検索
    - より意味を理解したシステム Q&Aシステム
  - 情報を分析したい
    - 企業, モノの評価を分析
  - 言語そのもの
    - 多言語に翻訳する際の事例

最近のトピック

- IBM Watsonの成功 (2011/2/16)
  - アメリカのクイズ番組 Jeopardy!
    - 「16世紀に初めて〇〇した人は?」
  - 歴代チャンピオン2名と勝負
  - 掛け金の総額で勝利!
- 基礎技術
  - 構文解析
  - 知識の構造化(UIMA)
  - Web上の文書
  - 統計的学習モデル

練習1

- 下記の質問に答えるためにどのような知識が必要?

  例: xp をsp2に上げたら無線LANが接続しなくなった. どうすれば良いですか?

言語処理の難しさ

- 表現の多様性
  - 同じことを表すことに多様な表現がある
    - 例) 「彼は鍵でドアをあけた」「ドアを鍵で彼はあけた」
- 文脈による意味の多様性
  - 同じ言葉でも文脈で意味が違う
  - 「僕はうなぎだ」 会話の一部として成立
  - “Time flies like an arrow.” 複数の意味
- 意味記述の不完全さ
  - 意味をどうかけばよいか?
    - 例) 「直線」, 「椅子」どう記述する?

→文字列による完全一致ではうまく処理できない

  - 計算能力や記憶量が進歩しても状況は変わらない

練習2

- 行く通りかの解釈を説明せよ
  - AさんとBさんは高校時代からの親友です
  - 黒い目の美しい女の子

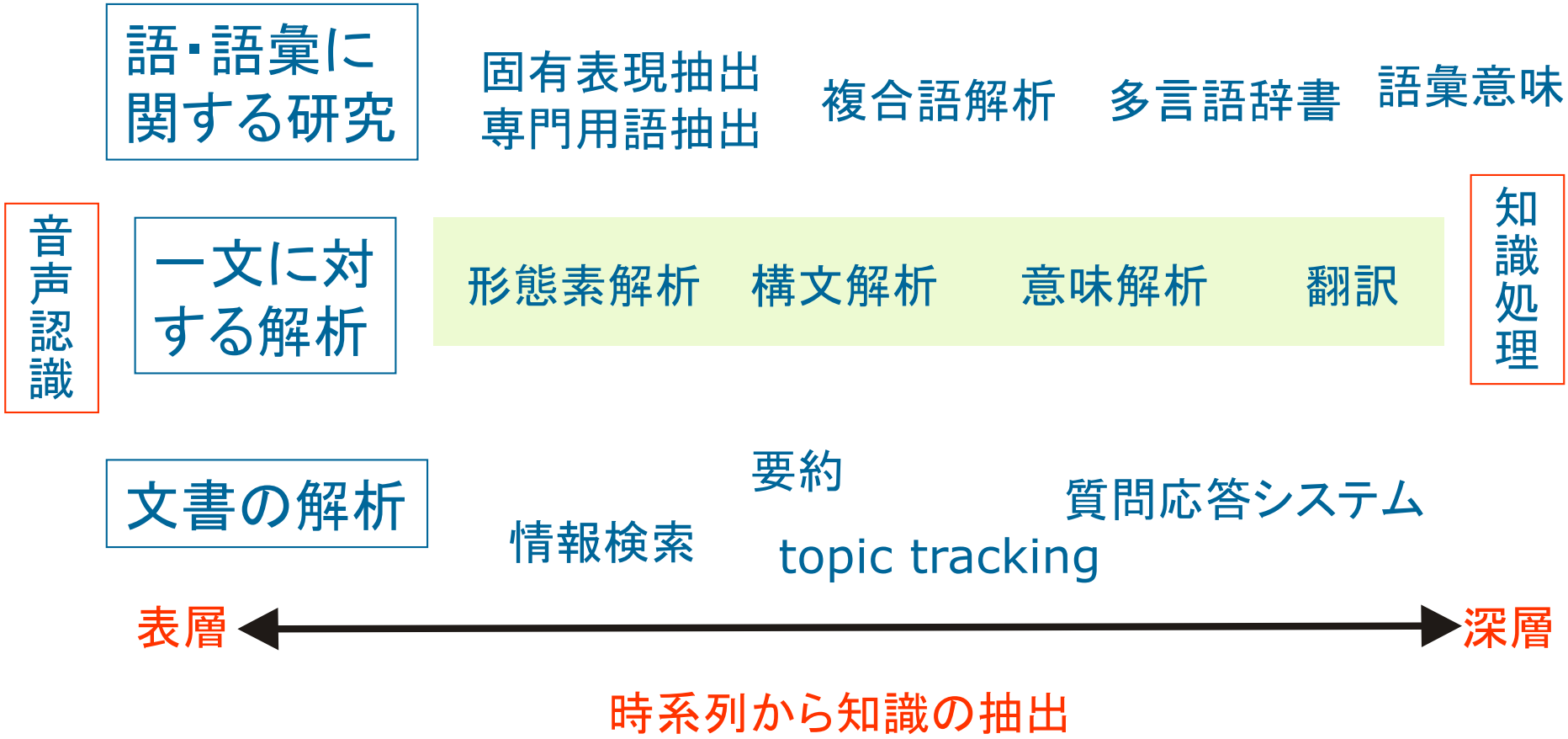
言語研究の流れ

- 言語そのものの研究
  - N. Chomsky 1957年
  - 言語の数学的なモデルと限界を提示
    →今日の言語モデルの基本
- 人間との結びつきを重視した研究
  - 認知心理学 1950年頃　スタート
- 知識と言語
  - 人工知能の研究 1960年　から
- コンピュータ上の処理
  - 機械翻訳(W. Weaver, A. D. Booth)
  - 1947年　英仏の翻訳システム

言語学
心理学
哲学

工学

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
情報検索
topic tracking
質問応答システム
表層
深層
時系列から知識の抽出

講義で扱う対象

- 言語の理論
  - 言語の入れ子構造を捉える数学的モデル
  - 言語のもつ制約を記述するモデル
- コンピュータ上の処理
  - 言語処理の各モデル
    - 形態素解析，構文解析，機械翻訳，データマイニング
      情報検索 (google)，質問応答
  - 言語資源と体系化
    - 言い換え，語の知識の構築

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

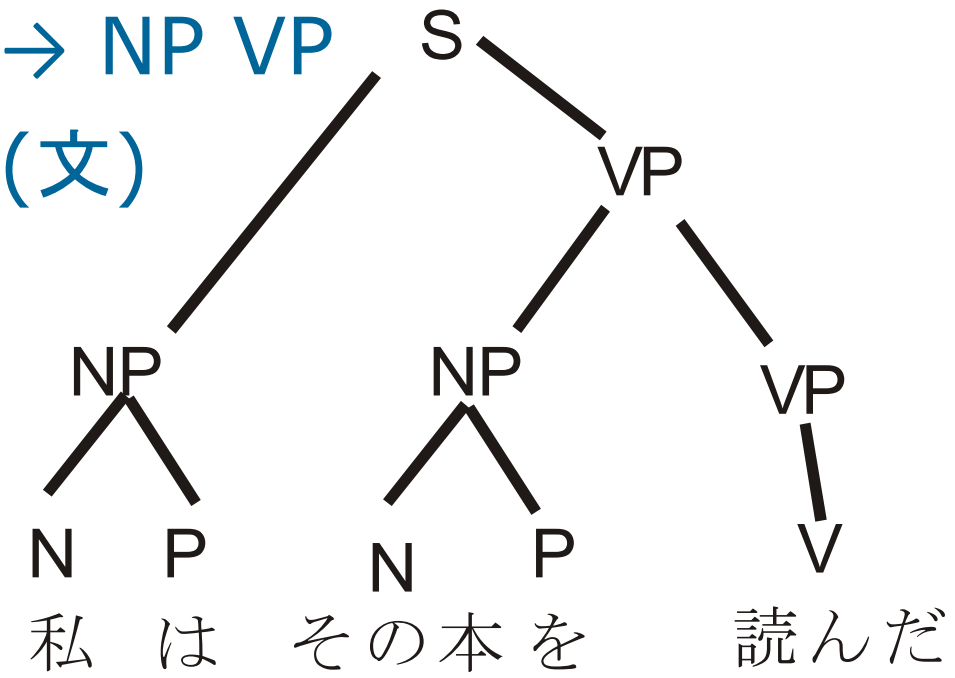
    刑事は　彼がその犯人であることを　知った

    (主語)(助詞)（目的語)(助詞)（述語)

形式化

- 形式文法 formal grammar
  - 非終端記号　NP (名詞句), VP (動詞句). .
  - 終端記号　私, は, 彼, 本, その
  - 生成規則　S → NP VP
  - 初期記号　S (文)

構文木
構文解析



練習3

- 先ほどの例の構文木を作成せよ
  - 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

形式文法

- 形式文法の定義 $G=\langle V_N, V_T, P, \sigma \rangle$
  - 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
  - 初期記号の集合: σ $\omega \in V_N$

  文の集合L(G)は

$$L(G) = \{w | w \in V_T^*, \sigma \overset{*}{\Rightarrow} w\}$$

  w はV_T からなる文字列

  V_T^*はV_Tの要素の任意の長さの文字列