自然言語処理2014(3回目)
Natural Language Processing 2014

東京工科大学
コンピュータサイエンス学部
亀田弘之

レポート課題1

- 自然言語処理システムの例をWeb等で1つ探し、その紹介文を作成しなさい。具体的には、システム名、システムの概要、システムの主要機能を簡単に説明してください。可能ならば図や写真を添えてください。A41枚～2枚程度でOKです。
- 提出日は次回平成26年9月29日（月）とします。A4レポート用紙を使い、必ず表紙を付けてください。

レポート課題2

1. “形態素”の言語学的定義を調べよ。
2. Winchaに関し以下のことを行え.
   - ① インストールする.
   - ② 次の例文を形態解析する.
   - ③ 形態素解析結果について，問題点があれば指摘する.

レポート回収します！

レポート課題No.2

1. “形態素”の言語学的定義を調べよ。
2. Winchaに関し以下のことを行え.
   - ① インストールする.
   - ② 次の例文を形態解析する.
   - ③ 形態素解析結果について，問題点があれば指摘する.

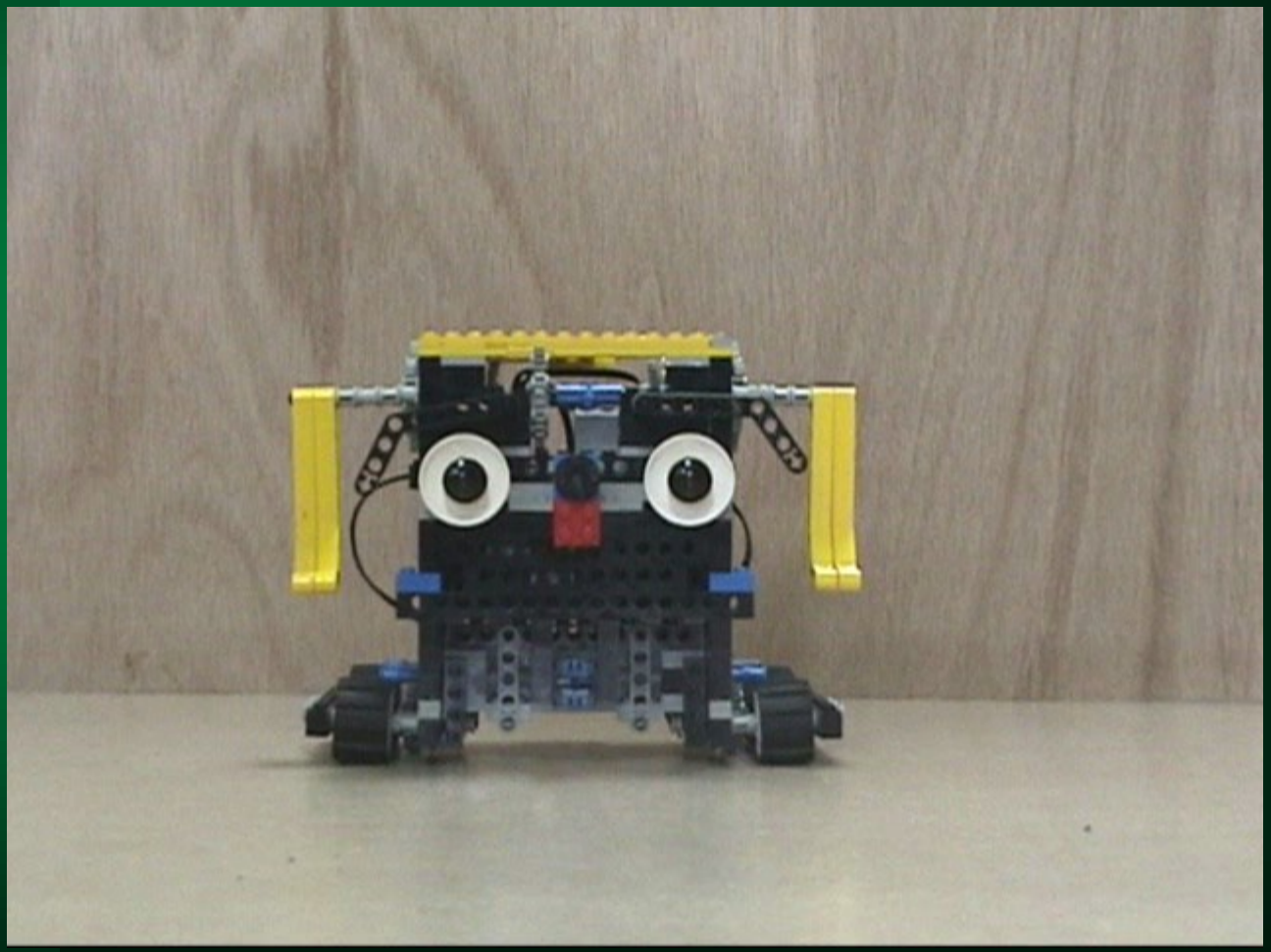
前回までの復習

- 言語 (languages)
  - 自然言語 (natural languages)
    - 文字言語 (written languages)
    - 音声言語 (spoken languages)
    - 視覚言語 (visual languages)
  - 人工言語 (artificial languages)
    - Programming languages
      - 手続き型言語・関数型言語・論理型言語
      - オブジェクト指向型言語・アスペクト指向型言語　など

前回までの復習

- 言語 (languages)
  - 自然言語 (natural languages)
    - 文字言語 (written languages)
    - 音声言語 (spoken languages)
    - 視覚言語 (visual languages)
  - 人工言語 (artificial languages)
    - Programming languages
      - 手続き型言語・関数型言語・論理型言語
      - オブジェクト指向型言語・アスペクト指向型言語　など

対話ロボットのデモ



自然言語処理の概要

- 文字認識(character recognition)
- 形態素解析(morphological analysis)
- 統語解析（構文解析, syntactic analysss）
- 意味解析(semantic analysis)
- 談話解析(discourse analysis)
- 状況解析(stuation analysis)
- 世界解析（他者理解など）

要素技術の応用分野

- 文字認識　→　スキャナの高度化
- 形態素解析　→　データマイニング
  情報検索
- 統語解析（構文解析）→　機械翻訳
  音声対話
- 意味解析　→　（同上）
- 談話解析　→　（同上）

形態素とは

- 意味を持つ最小の単位。
- 例:
  - Beautiful　beauti-ful
  - Books　book-s
  - 美しい　　美し-い
  - 語
  - 単語（語）　foot - feet　tooth - teeth
  - Word lexeme

Chasen, Juman, MeCab

- 日本語を対象とする形態素解析の代表的なプログラム
- ChasenかMeCabをインストールして使ってみよう。
  - www.vector.co.jp を通じて公開されている。
  - WinCha というWindowsインタフェースのものもある。(今はメインテナンスされてない。)(授業ページからダウンロード可能)

Winchaのインストール例

- WinchaとはChasenを元にして作られた形態素解析プログラムである。
  - 講義のページからダウンロード可。
  - 自動解凍形式になっている。
  - 解凍の際はadministrator権限で実行すること（特に、Windows XP では）。

自由課題：ツールを使ってデータ処理してみよう！

- 各自、新聞記事、小説、ブログなどさまざまなジャンルのテキストに対して、形態素解析ツールを利用して形態素解析してみてください。
（後日、レポートとして提出してもらいます。今日のレポート課題は別のものです。）

実行例

- 入力：生活の質の向上と文化の発展に貢献する人材を育成する
- 出力：生活/の/質/の/向上/と/文化/の/発展/に/貢献/する/人材/を/育成/する

東京工科大学の「基本理念」より引用

レポート課題No3

- 形態素解析ツール（MeCab, Chasen, Jumanのどれか1つ）を利用して、学長挨拶(http://www.teu.ac.jp/gaiyou/006488.html)を形態素解析しなさい。
- 提出日は、次回の授業の時とします。
- 表紙等もいつも通りとします。

参考(学長挨拶本文)

東京工科大学は1986年に工学系単科大学としてスタートしました。以来、日本初のメディア学部の設置など、常に社会のニーズを読みながら進化を続け、現在では5学部と大学院を擁し、蒲田と八王子の2キャンパスを有する総合大学へと発展しています。

本学は、新しい大学です。伝統校のような長い歴史はありませんが、新しい大学には、時代に即応した革命や新たな試みに挑戦できる柔軟な態勢があります。2012年度も新しい試みを実行します。それが「教養学環」の設置です。これにより、今までの教養教育を改革・充実させ、社会で必要とされる社会人基礎力、東京工科大学の学生として身につけてほしい教養を学部横断教育として行っていきます。しっかりとした教養を備えるということは、発展著しい社会の変化に適応できる普遍的な知識を身につけること。これを「東京工科大学教養スタンダード」とし、その後の専門教育の充実や学生の就業力向上をめざします。

東京工科大学の教育の根幹にある考えは「実学主義」です。「実学主義」とは「実社会で役立つ専門的な知識や技術、加えてその基盤となる人間としての適応力を高めるための教育」です。これは単に学問を修めるためだけでなく、卒業後、一人ひとりがそれぞれの分野で学んできたことを活かして、社会で活躍できる実践力を磨くことです。そのために、入学から就職・進学まで、一貫したサポート体制でみなさんの夢の実現を応援します。ぜひ東京工科大学の門を叩いて、飛び込んで来てください。

今日の後半に移りましょう！

構文解析（統語解析）

- NLPの中心的話題です。

まずは、背景にある理論から

- 言語理論 (Theory of Languages)
  - 処理対象そのものを知る
- 論理学 (Logic)
  - 処理のための理論(NLP in term of logic)
- プログラミング (Programming)
  - コンピュータで処理できるために

もう少し詳しく述べると．．．

- 言語理論
  - 形式言語(言語と文法、文脈自由文法)

    Formal languages (language & grammar, context-free grammar)
- 論理学
  - 述語論理（推論、レゾリューション法）

    Predicate logic (inference/reasoning, resolution method)
- プログラミング
  - 論理型プログラミング(Prolog)

    Logic programming

学習目標

- 簡単な構文解析プログラムを自力で作成できる。
  - 処理対象言語：日本語と英語
  - 使用プログラミング言語：Prolog

Learning goal
- to be able to design and implement a simple syntactic analyzer (parser) by oneself.

準備

- 次の文の構造を分析してみよう。

Tom broke the cup.

文法

- G = { Vn, Vt, σ, P }
  - Vn : 非終端記号
  - Vt : 終端記号
  - σ : 開始記号
  - P : 書き換え規則

- G = {Vn, Vt, σ, P}
  - Vn ={S, NP, VP, PrpN, V, Det, N}
  - Vt = {Tom, broke, the, cup}
  - σ = S
  - P={S→NP VP, NP →PrpN, VP →V NP, NP → Det N, PrpN → Tom, V →broke, Det →the, N→cup}

Prologの導入

- プログラミング言語としてはJavaやCでもいいのですが、本講義ではプログラミングの負担を減らすため、また、CS学部の学生ならば知っておくべき言語であるという理由から、Prologを用います。
- 元気な人は、ML, Ocaml, Haskell, Lisp, Python などでプログラミングしてください。

Prologの動作を理解する。

- 黒板で説明します。

予習問題

- 英文

  Tom broke the big cup.

  に対して以下のことを行え。

1. 統語構造（構文構造）を分析せよ。
2. この文を処理するための文法G_1を書け。
3. G_1を基にPrologプログラムを作成せよ。

次回以降、練習をします。

- PrologでNLPシステムを書けるようになりましょう！ みんな必ず書けるようになります。