自然言語処理2014　第2回目

東京工科大学
コンピュータサイエンス学部
亀田弘之

Natural Language Processing 2014

Tokyo University of Technology
School of Computer Science
Prof. Hiroyuki Kameda

レポート課題1（授業の最後に提出）

- 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。

まずは、復習から

Let’s remember what we studied last week.

自然言語処理とは（復習）

- 自然言語処理
  ＝．．．が　．．．を　．．．する

自然言語処理とは（復習）

- 自然言語処理
  ＝ 機械が自然言語を処理する

（注）本講義では、このように限定した範囲を扱う。

自然言語処理とは（復習）

- 自然言語処理
  ＝ 機械が日本語や英語を理解する

自然言語処理とは（復習）

- 自然言語処理
  ＝ 機械が日本語や英語を理解する

どうやって？
How?

自然言語処理の概要

- 形態素解析 (morphological analysis)
- 統語解析（構文解析, syntactic analysis）
- 意味解析 (semantic analysis)
- 談話解析 (discourse analysis)
- 状況解析 (situation analysis)
- etc.

今日の学習目標

- “形態素”という概念を知る。
- 形態素解析に慣れ，自分で解析できる。
- 形態素解析システムについて知る。
- 形態素解析手法の基本的考え方の例を知る。

形態素解析

- 形態素
- 形態素解析

形態素解析

- 入力：　メロスが走る
- 出力：　メロス(名詞)/が(助詞)/走る(動詞)

  - 処理方法は？
    - 手法
    - アルゴリズム
    - プログラミング言語
  - 必要な知識は？
  - 評価方法は？
  - ツールは？

例：

- Tom broke the cup with the hammer.
- Tom brach die Tasse mit der Hammer.
  (Tom hat die Tasse mit der Hammer gebrochen.)
- Tom a cassé la tasse avec le marteau.

- Tom broke the cup with a state of the art hammer.

形態素解析

- 入力：　メロスが走る
- 出力：　メロス(名詞)/が(助詞)/走る(動詞)

  - **処理方法**は？
    - 手法
    - アルゴリズム
    - プログラミング言語
  - 必要な知識は？
  - 評価方法は？
  - ツールは？

メロスが走る

メロスが走る

メロス / が / 走る

処理方法

- 文字種法
- 最長一致法
- 文節数最少法
- 接続表を利用する方法
- 遷移確率を用いる方法　etc.

例文1：

読売新聞社が２～４日に実施した全国世論調査（電話方式）で、鳩山内閣の支持率は７１％、不支持率は２１％だった。

2009年10月4日　Yahoo Japan ニュース（読売新聞）より

例文2

- 北極圏で今春、観測史上最大のオゾン層破壊が起きていたことが、国立環境研究所（茨城県つくば市）など９か国の国際研究チームの分析で分かった。

2011年10月3日　Yahoo Japan ニュース（読売新聞）より

例文3

- 臨時国会の召集の遅れは、野田政権が最重要視する社会保障・税一体改革にも影を落としている。

2012年10月14日　Yahoo Japan ニュース（読売新聞）より

例文4

強い台風24号は7日、勢力を保ったまま沖縄本島や鹿児島県・奄美地方に接近する見込みだ。気象庁は、両地域が7日昼過ぎから風速25メートル以上の暴風域に入る恐れがあるとして、暴風や高波への警戒を呼びかけている。

(2013年10月7日，朝日新聞デジタル)

例文5

御嶽山の噴火から一夜明けた28日、山中に取り残された登山者らの救助は難航した。自衛隊や警察、消防などは長野県側だけで21人を救助。ただ、灰に埋もれ、心肺停止した人々を確認しながら、強い硫黄の臭気に阻まれ、多くの搬送が断念に追い込まれた。

(2014年9月29日　朝日新聞デジタル)

例文6　くるまでまつ

- 他の例:

花子は太郎と次郎をずっと
くるまで待ち続けていた。

曖昧性(Ambiguity)

自由課題1

- 日本語を対象とする形態素解析プログラム（形態素解析器）の発展史を調べ、現状と今後の問題点を考えよ。

参考情報：
Juman, Chasen, すもも, Kobako etc.

自由課題2

- 英語を対象とする形態素解析プログラムとして、何があるが調べなさい。

  ヒント：tagger programs

ヒント

- 最長一致
  - KAKASI
- Bi-gramマルコフモデル
  - JUMAN, MeCab
- 可変長マルコフモデル
  - ChaSen
- 未知語処理（綴り・品詞・意味・用法）
  - UWAS-I
- 未知統語規則処理
  - Progol

提出方法

1. 書式：
   - A4レポート用紙
   - 表紙を付けること（日付，氏名，学籍番号）
2. 提出日：
   - 平成24年10月21日（月）授業時間中
3. 提出場所：
   - 教室

問題：次の文を形態素解析せよ。

- 地球温暖化を防ぐために世界の国々がつくった京都議定書で、日本は二酸化炭素(CO_2)などの温室効果ガスを2008から12年度に1990年に比べて6%減らすことを約束しています。
（朝日小学生新聞2007年10月12日　より）

国語の問題として考えてみよ。

形態素解析結果

（続き）

- 自然言語は人間相互の意思疎通のための道具であり、これを機械により処理することが出来れば、社会的意義は極めて大きなものがある。

文字種法

- 文字種類の変わり目を単語の切れ目とする考え方。切り出した後、微修正が必要。

- 例:
  文字種類 / の / 変 / わり / 目 / を / 単語 / の / 切 / れ / 目 / とする / 考 / え / 方 / 。 / 切 / り / 出 / した / 後 / 、 / 微修正 / が / 必要 / 。

例7

- 地球温暖化 / を / 防 /ぐために / 世界 / の / 国々 / がつくった / 京都議定書 / で / 、 / 日本 / は / 二酸化炭素 / (/ CO_2 /) / などの / 温室効果 / ガス / を / 2008 / から / 12 / 年度 / に / 1990 / 年 / に / 比 / べて / 6 / % / 減 / らすことを / 約束 / しています / 。

日本語における字種

- 区切り記号（句読点、特殊記号）
- 漢字、片仮名、アルファベット（英文字）
- 数字
- 平仮名

形態素解析結果の第一次近似解を得るヒューリスティック(heuristic)

- 平仮名から他の文字種への変わり目
- 区切り記号の前後
- 非平仮名列から数字列への変わり目
- 数字列から非平仮名列への変わり目

  - 以後、さらに精度を高める。
    何をすればいいのか？　考えてみよう。

前記ヒューリスティック適用例

- 文字種類の変わり目を単語の切れ目とする考え方。切り出した後、微修正が必要。

- 文字種類の/変わり/目を/単語の/切れ/目とする/考え/方/。/切り/出した/後/、/微修正が/必要/。/
- 文字種類 / の / 変 / わり / 目 / を / 単語 / の / 切 / れ / 目 / とする / 考 / え / 方 / 。/ 切 / り / 出 / した / 後 / 、 / 微修正 / が / 必要 / 。

問題:文字種法の長所・短所

- 長所は、. . .

- 短所は、. . .

問題：文字種法の改良案を考えよ。

- （例に基づいて考えること）

問題:文字種法の適用分野はあるか？あるとすれば何？

最長一致法

- 処理対象文字列の先頭から始まる単語のうち、文字列長が最大のものを優先的に単語候補とする方法。

例

- アルプスのやまは美しい
- アルプスの少女は美しい

- 単語辞書:
  - アルプス
  - 少女
  - やま
  - のやま
  - 美しい
  - は
  - の

問題：必要な単語辞書を作成せよ。

- ユク河ノナガレハ、絶エズシテ、シカモモトノ水ニアラズ。

- 単語辞書：
  - 河
  - 水
  - ユク
  - ノ
  - ナガレ
  - 絶エズシテ
  - シカモ
  - モトノ
  - ニアラズ

（出典：『方丈記　徒然草　正法眼蔵随聞記　歎異抄』 日本古典文学全集　昭和46年　小学館）

問題:最長一致法の長所・短所

- 長所

- 短所

文節数最少法

- 文節数が最も少なくなる切り方を解とする方法。

例:くるまでまつ

- 車で　待つ　　（2）
- 車で　松　　（2）
- 来るまで　待つ（2）
- 来るまで　松　（2）
- 繰るまで　待つ（2）
- 繰るまで　松　（2）
- 狂まで　　待つ（2）
- 狂まで　　松　（2）
- 来る　間で　待つ　　（3）
- 来る　間で　松（3）etc.

接続表を利用する方法

遷移確率を用いる方法

- n-gram

各種ツール

- Chasen (WinCha)
- Kobako-J
- XMLEDITOR.NET
- GoTagger　など
  （この他にもいくつかあります。
  調べてみなさい。）

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.

例文

（授業の時に提示する．）

レポート課題1（提出お願いします！）

- 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。