プログラミング応用
http://bit.ly/kosen02

Week14@後期(week29 in 2016)
2016/12/22

お知らせ

- 補講は1/26(木; 水曜日課)の7-8限に行います
  (ホームルーム終了後、演習室に来てください)

(復習)言語の種類

1. 自然言語
   - 人間が普段情報伝達に用いる言語
     特徴:”曖昧性”を含む
     例) I want to eat somewhere near the sea.
2. 人工言語
   - 人間が何かしらの目的のために作り出した言語
     例) C言語、シェルスクリプト(Bash)
     楽譜なども人工言語
     特徴:”曖昧性”がない

(復習) 計算機による人工言語解析

1. 字句解析
   - 入力言語を「字句(トークン)」の並びに分割する処理
   - 字句: 受理する文字列を正規表現で定義(lex)
2. 構文解析
   - 字句解析した言語から構文木を構築する処理
   - 構文木構築: 文脈自由文法で規則を定義(yacc)
3. 意味解析
   - 字句解析、構文解析した結果に曖昧性がある場合にもっともらしい意味を推定
   - 人工言語ではほぼ不要

本日の内容

- 講義(自然言語処理)
  - 自然言語処理の概要
    - 自然言語の特徴
    - 自然言語処理の例
- 演習
  - 既存ツール(MeCab)を用いた形態素解析

自然言語の特徴

- 自然言語には字句解析、構文解析、意味解析すべての段階で曖昧性(ambiguity)が存在
  - 例1:
    「ニワニワニワニワトリガイル」という音声を文字起こしするプログラムを作成するには?
    → 庭には2羽鶏がいる?
    → 2話には2羽鶏がいる?
  - 例2
    I eat somewhere near the sea.
    → 海の近くのどこかを食べる?
    → 海の近くのどこかで食べる?

自然言語処理の例

1. 字句解析
   - 英語: 単語分割+品詞ダグ付け
   - 日本語: 形態素解析+品詞タグ付け
2. 構文解析
   - 句構造解析(Phrase structure analysis)
   - 係り受け解析(Dependency analysis)
3. 意味解析
   - 語義曖昧性解消、極性判定

プログラミング応用では自然言語を扱うために必要な処理を概観し、既存のツールを用いて言語処理をする

自然言語での字句解析

- 英語
  I flied to San Francisco from Tokyo.
  → 字句解析はスペースで分割するだけでは不十分(San Francisco等が分割されてしまう)

- 日本語
  私はペンを持っています。
  →そもそもスペースで区切られておらず字句解析自体が難しい

単語分割(word segmentation)

- 文を単語に分割する(主に英語で必要)
  - スペースで区切られた文で必要

形態素解析

- 入力文を形態素という基本単位に分割する
  - 形態素: 意味を持つ最小単位
  - 日本語/中国語などスペース区切りされていない言語でよく使われる処理

私 は ペン を 持って います 。

品詞タグ付け

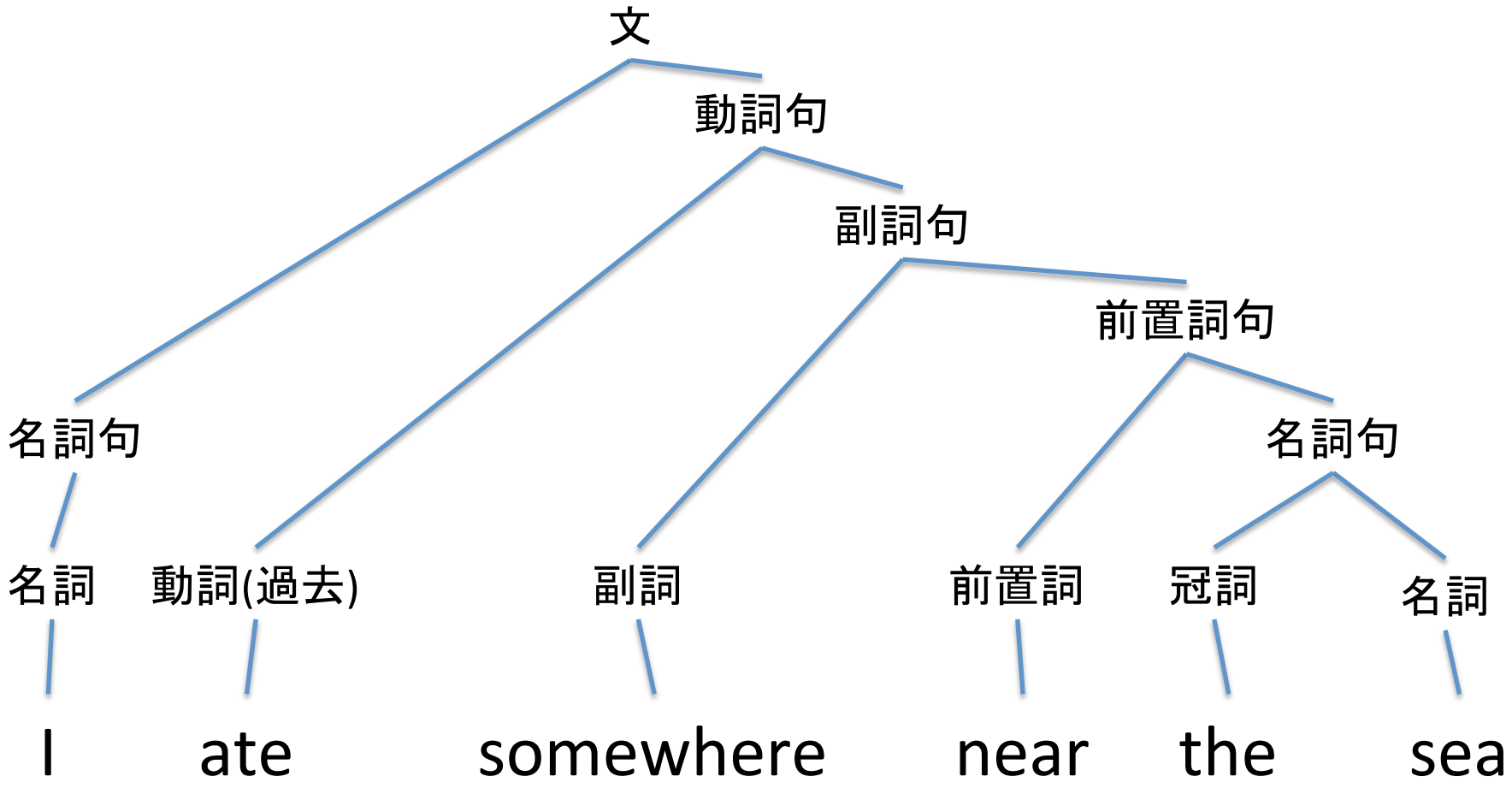
- 単語や形態素に”品詞”情報を付与する

自然言語処理の例

1. 字句解析
   - 英語: 単語分割+品詞ダグ付け
   - 日本語: 形態素解析+品詞タグ付け
2. 構文解析
   - 句構造解析(Phrase structure analysis)
   - 係り受け解析(Dependency analysis)
3. 意味解析
   - 語義曖昧性解消、極性判定

句構造解析

- 単語間の文法関係を木構造で表現する



係り受け解析

- 文節間の係り受け関係を解析する

自然言語処理の例

1. 字句解析
   - 英語: 単語分割+品詞ダグ付け
   - 日本語: 形態素解析+品詞タグ付け
2. 構文解析
   - 句構造解析(Phrase structure analysis)
   - 係り受け解析(Dependency analysis)
3. 意味解析
   - 語義曖昧性解消、極性判定

語義曖昧性解消

- 複数の語義(単語の意味)がある場合に文脈に応じて適切な語義を選ぶ

I went to the bank in the university.

銀行

I went to the bank of a river

川岸

極性判定

- 文がPositiveな意味を持つか、Negativeな意味を持つか推定する

ステーキが厚い

パソコンが厚い

演習概要

- 本演習では形態素解析のみ扱う
- 形態素解析ソフト「MeCab」とUNIXコマンドを用いた言語解析演習

演習1-x: MeCabの使い方を確認
演習2-x: MeCabでファイルを読み込む方法の確認
演習3-x: MeCabとUNIXコマンドで
「吾輩は猫である」を解析

演習1-1

- 以下のコマンドを入力しMeCabを起動しなさい

  $ ~ishigaki/mecab

- 「すももももももものうち」と入力しエンターキーを押し、名詞の「もも」が何回出現したか確認しなさい。

- 形態素解析せずにgrepコマンドで「もも」の出現回数を数えるとどのような問題があるか考察しなさい
  (提出不要)

演習1-2(MeCabの終了)

- Ctrl+Cを押してMeCabを終了しなさい。

演習1-3

- MeCabを再度起動し他の文も解析してみましょう。(提出不要)

演習2-1

- テキストエディタgeditで、sample.txtというファイルに以下の内容を書き込みなさい。

```
これはMeCabのテストです。
このファイルを読み込んでいます。
```

演習2-2(ファイルの解析)

- MeCabでファイルを解析するには、演習2-1の用に1行に1文を記述し、以下のようにMeCabを実行します。

  $ ~ishigaki/mecab < sample.txt

  sample.txtの内容が正しく形態素解析されることを確認しなさい。
  (提出不要)

演習2-3(解析結果の書き出し)

- 以下のコマンドで、sample.txtの解析結果をsample.mecabというファイルに書き出す事ができる。

  $ ~ishigaki/mecab < sample.txt > sample.mecab

  (提出不要)

演習3-1

- 「吾輩は猫である」の全文データを形態素解析し、neko.mecabというファイル名で保存しなさい

- 「吾輩は猫である」の全文データは以下にある。手元にない人はcpコマンド等でコピーして使うと良い
  ~ishigaki/neko.txt

演習3-2

- grepコマンドを用いてneko.mecabのうち名詞を含む行を抽出し、neko.mecab.nounというファイル名で保存しなさい

演習3-3

- sortコマンドとuniqコマンドを用いて neko.mecab.nounに含まれる行をカウントし、neko.mecab.noun.uniqというファイル名で保存しなさい

演習3-4

- sortコマンドを用いて「吾輩は猫である」に出現する名詞を出現回数順に並べ替え、neko.mecab.noun.uniq.sortというファイル名で保存しなさい

(早く終わった人)演習4

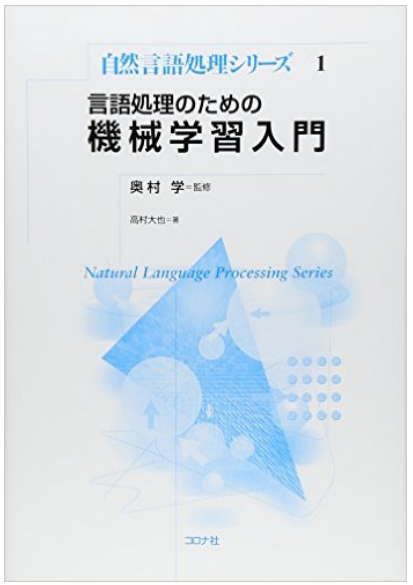
- 「吾輩は猫である」の名詞、固有名詞、動詞、助詞、記号の出現回数をそれぞれUNIXコマンドを組み合わせ、出現回数が多い順に上位２０件を表示しなさい。

  コマンドは1行で記述すること。

  (macab, uniq, sort, wcなどをパイプ | でつなぐと良い。)

さらに深く学ぶ人向け

自然言語処理の基礎
奥村学



言語処理のための機械学習入門
高村大也

次回

- これまでは”文”を解析する方法を学習しましたが、次回は”文書”を扱う方法を扱います

試験

- 2/2 午後(予定)
- マシン内の資料、プログラムのみ参照可