自然言語処理2014 No.11

東京工科大学
コンピュータサイエンス学部
亀田弘之

今までの振り返り

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）

NLPのプログラムを書いてみよう！

I. Prologのインストール

II. プログラム作成手順

1. 解析対象テキストの収集
2. IC分析（直接構成素分析，**Immediate Constitutent Analysis**）
3. 形式文法の設定
4. Prolog形式への書き換え
5. NLPプログラムの実行(その１)
6. 構文木を出力するプログラムへの拡張
7. NLPプログラムの実行(その２)

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

文法の例

- 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}

第5回目の資料より

3. 形式文法の設定

- 開始記号＝s
- 終端記号={ tom, broke, the, cup }
- 非終端記号={ s, n, v, d, np, vp }
- 書き換え規則={

  s → np + vp.　np → n.　np→d+n.　vp→v+np.

  n→tom.　n→cup.　v→broke.　d→the.

}

第5回目の資料より

4. Prolog形式への書き換え

s :- np,　vp.

np :- n.

np:-d, n.

vp:-v, np.

n:-tom.

n:-cup.

v:-broke.

d:-the.

第5回目の資料より

Prog1.pl

```
s(A,C):-n(A,B),vp(B,C).
vp(A,C):-v(A,B),np(B,C).
np(A,C):-d(A,B),n(B,C).
n([tom|T],T).
n([cup|T],T).
v([broke|T],T).
d([the|T],T).
```

第5回目の資料より

Prog2.pl

```
s(A,C,s(_n,_vp)):- n(A,B,_n),vp(B,C,_vp).
vp(A,C,vp(_v,_np)):- v(A,B,_v),np(B,C,_np).
np(A,C,np(_d,_n)):- d(A,B,_d),n(B,C,_n).

n([tom|T],T,n(tom)).
n([cup|T],T,n(cup)).
v([broke|T],T,v(broke)).
d([the|T],T,d(the)).
```

第5回目の資料より

3. 形式文法の設定

- 開始記号＝s
- 終端記号={ tom, broke, the, cup }
- 非終端記号={ s, n, v, d, np, vp }
- 書き換え規則={

  s → np + vp.　np → n.　np→d+n.　vp→v+np.

  n→tom.　n→cup.　v→broke.　d→the.

}

まずは文を集めてこなければ、上記の情報を作れない！

コーパス

- コーパス(Corpus)は、コンピュータの発達とともに、計算機可読なデータを容易に作成・収集することができるようになったことがその背景にある。現在ではコーパス言語学などの学問もある。

コーパスの定義

- 定義：まとまりのある文書データベースのこと。
  - シェイクスピアコーパス
  - 夏目漱石コーパス etc.

  現在では、大規模なテキストデータベースのことをコーパスと呼ぶこともある。なお、近年は多くのものにタグが付けられている。

コーパスの例

- Brown Corpus(米国の書籍・新聞・雑誌)
- LOB Corpus(英国の書籍・新聞・雑誌)
- British National Corpus(BNC)
  （英国英語、多様なジャンル）
- Bank of English
- Penn Treebank(Wall Street Journal)
- EDRコーパス（日本語）
- 日本語話し言葉コーパス
- 日英新聞記事対応付けコーパス　など

言語資料関係のサイト

- LDC(www.ldc.upenn.edu)
- ELRA(www.elra.info)
- GSK(言語資源協会, www.gsk.or.jp)
- RSC(音声資源コンソーシアム, research.nii.ac.jp/src/)

- 言語情報処理ポータル: nlp.kuee.kyoto-u.ac.jp/NLP_Portal/

- Gutenberg Project
- 青空文庫
- など

自然言語の応用

- 情報検索

情報検索

- Information Retrieval(IR)はWebの発展に伴い、ますますその重要性を増している。多くのWebは自然言語で書かれており、また、自然言語による検索は多くの人にとって便利である。

参考文献: Spidering Hacks (Python言語)

機能語と内容語

- 自然言語は人間相互の意思疎通のための道具であり、それをコンピュータにより処理することは社会的に意義のあることである。

問：どれが機能語でどれが内容語か？

検索の方式

- ディレクトリ方式
- キーワード方式

検索の方式

	手作業分類	自動分類
ディレクトリ方式		
キーワード方式		

（注）自動分類の際には，データ収集も自動的に行われていることが多い．自動収集用ソフトウェアを，crawlerとか検索ロボットなどと呼ぶ．

今日の課題：検索方式について

- どのような検索があり得るか？
  - 画像をキーとする検索
  - 画像を検索対象とする検索
  - 画像の他に音楽データ・楽譜などもあり得る。

  - もっと他には？
    - まだありますよね！考えてみてください。
    - さらに“それ”と言語との関係も考えてみてください。

自由課題　提案してみよう！

キーワードの見つけ方

定義：　キーワード

　　＝　そのページ・文章等で重要な用語

疑問：　どうやって見つけるのだろうか？

キーワードの見つけ方

定義：　キーワード

＝　そのページ・文章等で重要な用語

疑問：　どうやって見つけるのだろうか？

その1つに，tf-idf法がある。

語の重要度の計算法

- tf･idf法

$$tf \cdot idf = tf \times \left(\log \frac{N}{df} + 1 \right)$$

具体例で理解しよう！

tf･idf法の考え方(1)

文書	キーワード
Doc1	言語　コンピュータ　問題
Doc2	コンピュータ　問題　情報
Doc3	言語　問題　情報
Doc4	問題　情報

（注）キーワードを「索引語」ということもある.

tf･idf法の考え方(2)

キーワード	文　　書
言語	Doc1 Doc3
コンピュータ	Doc1 Doc2
問題	Doc1 Doc2 Doc3 Doc4
情報	Doc2 Doc3 Doc4

tf･idf法の考え方(3)

TF	Doc1	Doc2	Doc3	Doc4	IDF
言語	2	0	1	0	2
コンピュータ	1	1	0	0	2
問題	2	2	3	1	1
情報	0	1	2	1	1.3

$$IDF = \frac{\text{文書総数}}{\text{語が出現する文書の総数}}$$

tf･idf法の考え方(4)

TF･IDF	Doc1	Doc2	Doc3	Doc4
言語	4	0	2	0
コンピュータ	2	2	0	0
問題	2	2	3	1
情報	0	1.3	2.6	1.3

検索モデル

- ブーリアンモデル(Boolean model)
- ベクトル空間モデル

ブーリアンモデル

- 検索式1＝　コンピュータ　and　マック
- 検索式2＝　not マック　and ハンバーガ

ベクトル空間モデル

- D1, D2, …, Dn：「n個の文書」
- これらの文書全体に「m個の索引語」

n×mの行列で表現できる！

	言語	PC	Java	HTML
D1	2	0	1	0
D2	1	2	3	5
D3	0	2	8	3
D4	5	7	2	9

コサイン尺度

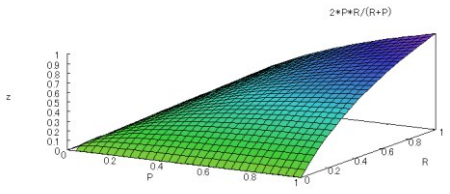
$$\cos(q, D_j) = \frac{\sum_{k=1}^{m} q_k d_{kj}}{\sqrt{\sum_{k=1}^{m} q_k^2} \sqrt{\sum_{k=1}^{m} d_{kj}^2}}$$

例：

- 教科書p.143参照

検索結果の評価式

- 適合度(精度, Precision)
- 再現率(Recall)

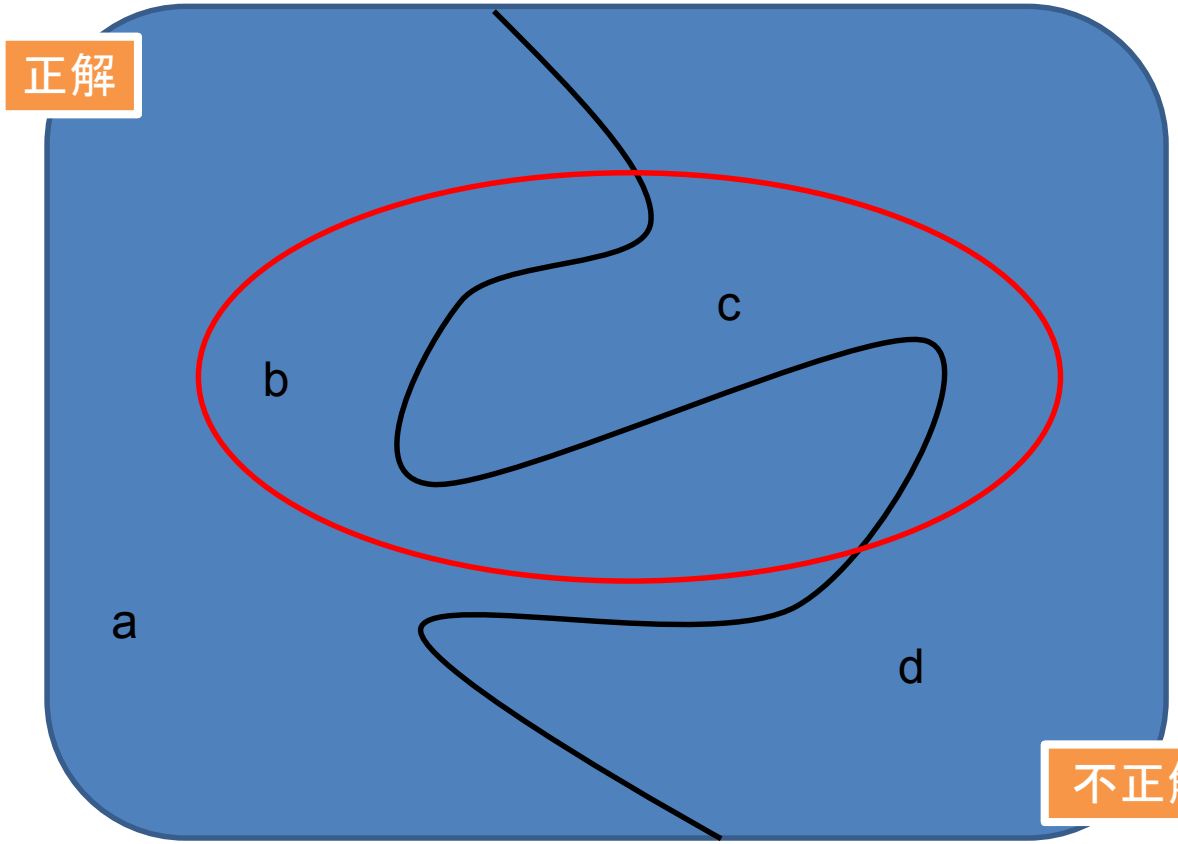


PとRの調和平均F値をとる。

$$F = \frac{2PR}{P+R}$$

（参考）右上の図は，Maximaでのコマンド plot3d(2*P*R/(P+R),[P,0,1],[R,0,1]);により作成.

検索評価の練習



$$P = \frac{b}{b+c} \times 100$$

$$R = \frac{b}{a+b} \times 100$$

総合演習1

1. 次の文章を形態素解析しなさい。
2. 統語構造を解析しなさい。
3. 意味を理解し、その内容を表現することのできるデータ構造（知識表現）を考えなさい。

<デジタル教科書>政府が「検討」 端末数万円、一部自己負担も

学校教育法で認められていない「デジタル教科書」を、２０１６年度にも解禁する方向で政府が検討を始めることが１２日、分かった。教育現場での実証研究を進めるとともに、教科書検定制度のあり方などの課題を１４年度までに整理する。政府のIT総合戦略本部（本部長・安倍晋三首相）が今月中にとりまとめる規制改革アクションプランに盛り込む。(毎日新聞)

前大統領派と治安部隊が銃撃戦＝きょう挙国一致政権発表―チュニジア

時事通信 2011年1月17日(月)7時10分配信

【カイロ時事】ベンアリ政権崩壊後の混乱が続くチュニジアの首都チュニス郊外の大統領府周辺で16日夜、前大統領派の部隊と治安部隊による激しい銃撃戦が展開された。一方、メバザア暫定大統領に組閣を命じられたガンヌーシ首相は「あす（17日）チュニジア史の新たなページを開くことになる新政権を発表する」との声明を出した。

AFP通信によれば、大統領警護部隊が籠城していた大統領府に軍部隊が攻撃を仕掛け、銃撃戦に発展した。治安部隊はまた、首都中心部の内務省付近の建物に銃を持って隠れていた2人を射殺。野党本部近くでも銃撃戦が発生した。

（Yahooより引用）

ソニー「3DSの好調に勇気づけられている」…ゲーム専用機の市場はある

インサイド2012年1月15日(日)15時21分配信

- 欧米では2月の発売を予定しているPlayStation Vita。日本国内ではスロースタートとなっていますが、ソニー・コンピュータエンタテインメント・ヨーロッパのJim Ryan社長兼CEOは業界紙MCVのインタビューに答え、3DSの好調には勇気づけられていると述べました。「ゲーム専用の携帯デバイスにはもはや市場は無いと言う人もいます。しかしクリスマスの3DSの好調な売上は勇気づけてくれるものです」Ryan氏は(3DSとVitaという)2つのゲーム機は市場を分け合う事が出来ると言います。また、スマートフォンとの競合については「我々がVitaで提供しようとしているクオリティ、没入感、リッチな体験は、どんなスマートフォンでも実現できないレベルのものです」

（Yahooより引用）

おわりに

- 自然言語処理の研究は盛んに行われていますが、まだまだ研究すべきものが残っています。
- みなさんも積極的に自然言語処理の研究にチャレンジしてください。
- 少なくとも、自然言語が人間社会で果たしている役割を考えれば、新しいアプリケーションのアイデアも湧くと思います。

定期試験について

- 過去問（後日配布）を良く勉強してください。
- 文法の作成とそれに基づくPrologプログラムの書き方を勉強してください。
- 新しく画期的な自然言語処理システムを考案してください。
- 基本的な用語の説明ができるようになってください。
- 持ち込み不可です。

（詳細は後日行います．）

基本的用語とは

- 言語
  定義と分類：自然言語，音声言語，視覚言語
- 自然言語処理
  - 形態素解析・統語解析・意味解析など
- 木構造
- 意味
  - 意味表現，知識表現
- コーパス　など

次回予告

- 文法獲得（統語規則獲得）について紹介。
- 帰納論理プログラミング(Inductive Logic Programming; ILP)の話をします。
- 予習等は特に必要ありませんが、興味のある人は、Web等を見てください。