情報抽出の基礎
―用語抽出―

東京大学　情報基盤センター
（情報理工学系研究科、情報学府兼担）
中川裕志

内容概略

- 情報抽出には抽出すべき情報の複雑さによっていろいろな方法が研究されている。
- ここでは、最も基礎になる抽出の単位として用語（＝コーパスにおける重要語）の抽出方法を説明する。
- 用語とは
  - term(用語、専門用語、術語）
  - terminology（学術用語、術語）
- コーパスからの用語抽出
  - 用語、連語
  - 統計的用語、コロケーション抽出
  - 構造と統計による抽出法
  - 用語生成

目的と抽出元のコーパス

- 機械翻訳用辞書
  - 新聞、ジャンル別コーパス
- 情報検索用キーワード抽出
  - 検索対象の文献DB
- 文書の索引語抽出
  - 索引を作るべき本
- ある学問、技術分野の用語抽出
  - その分野のコーパス

用語の定義付け

- 定義
  - 用語 term: 分野の概念を表す単語ないし複合語
  - 専門語彙 terminology: 分野の語彙。すなわちtermの集合
- 一般の語彙と専門用語の差
  - 現われる文書に偏りあり
  - 使用する人々（一般人 vs 専門家）
  - 用語の意味が一意的（円滑なコミュニケーションのため）
- ある分野(domain)における概念を表わす語彙
  - 歩留まり、横持ち、玉がけ、頭繋ぎ、腹くくり？？？
- 専門文書、コーパスにおける出現の仕方により定義付ける

termhood vs unithood

- unithood(単位性):複合語や連語（collocation）においてsyntagmatic（語順、構文構造、意味的関係などのこと）な関係が安定して用いられている度合い
- termhood(用語性):複合語や連語が、領域あるいは対象分野固有の概念と関連する度合い

自動用語抽出の枠組み

1. 分野の文書集合（コーパス）を形態素解析し品詞タグをつける
2. 用語にふさわしい単語の連続を品詞タグなどを参考にして抽出。これを用語候補とする。
3. 用語候補に用語らしさを反映するスコアをつける。
4. スコア順に並んだ候補から適当なものを用語として抽出（例えば、決められた数を選ぶ、など）

unithoodによる用語候補の抽出

- 抽出したいものが単語であるので文字N-gramは向かないが、単語抽出が難しい中国語やドイツ語では使わざるをえない。
- 日本語なら形態素解析して単語分割
- 各単語に品詞タグを付ける
- （専門）用語らしい品詞列（途中に特定の単語を含んでよい）を定め、それに一致する単語列を用語候補とする。
- どのような品詞列を選ぶべきか？
- さらに、専門用語としてはどのような品詞列、あるいは語構成があるのか？

用語の文法的構造

- 日本語の場合
  - 名詞+、　例：情報システム、井戸型ポテンシャル、チョムスキー階層
  - 名詞+「の」名詞+　、例：言語の分析
  - 形容詞　名詞+、　例：大域的（な）制御
  - 数詞　名詞、例：　3型言語
- 形容詞とは　イ形容詞：大きい

  ナ形容詞：絶対的

用語の文法的構造

- 英語の場合
  - →以下で | は or, A^+ はAの1回以上の繰り返し,Aは $A^?$の0 or 1回
  - 名詞$^+$、 例： computer network
  - 名詞$^+$“of”名詞$^+$ 、例： lack of stimulus
  - 名詞　前置詞　名詞、例：
  - 形容詞　名詞$^+$、 例： global data, balancing act
  - 数詞　名詞$^+$ 、例： first order logic
- まとめると
  - ((形容詞|名詞)* |(形容詞|名詞)*(名詞句) $^?$)(形容詞|名詞) *)名詞

用語の文法的構造

- フランス語の場合
  - 名詞“de”名詞、 例：assemblage de paquet, reseau de satellites
- まとめると、
  - 名詞　形容詞，名詞　名詞，名詞 de 名詞、名詞　前置詞　名詞

単名詞の用語らしさ(termhood)の定量的尺度

- dj: document set of domain, $D=\{d_1,\ldots,d_j,..d_{n(D)}\}$
- wj: word appeared in D, $W_D=\{w_1,\ldots,w_i,..\}$
- wij: word(wi) appeared in d_j
- f(wij)= d_j中のw_iの出現頻度
- g(wij)= 1 when w_i is in d_j,
  0 when w_i is not in d_j
- 文書での出現頻度に基づく2つの尺度:
  - average word freq in d_j = $If_{ij} = \frac{f(wij)}{\sum_i f(wij)}$

- ある単語 i が文書 j に特別な現れ方をするかどうか

$$Iij = \frac{f(wij)}{\sum_i f(wij)} - \frac{\sum_j f(wij)}{\sum_i \sum_j f(wij)}$$

- Similar idea: tf × idf

$$tf \times idf_{ij} = f(wij) \times \log\left(\frac{n(D)}{\sum_j g(wij)} + 1\right)$$

- 以上の2つは、ある文書d_jだけで特徴的に多く出現するw_{ij}を優先する考え方

用語性の計算法まとめ

- ✓ 以上の方法はいずれも termhood を測ろうとしたもの。まとめると、
  - ✓ ある文書に頻出する単語が用語
  - ✓ 限定された文書にだけ出現する単語が用語
  - ✓ 全文書の中である文書にだけ際立って頻出する単語がその文書の用語(or index term)
  - ✓ 全文書において出現分布に特徴のある単語が用語
  - ✓ Etc
- ✓ これらは全て文書集合における用語の性質による→文書空間　(document space based method)
  - ✓ 後に違う見方（語彙空間による見方）を紹介

複合語、連語（collocation）の unithood, termhood　文書空間法

- まず、安定して使われる用語かどうか(unithood)を調べることになる。
  - 基本単語(複合語でない単語)が偶然より統計的に有意に高い頻度で共起するかを調べる

  → Contingency matrix
  - これは文書空間における単語間の統計的性質を利用する方法

Contingency Matix

- 二つの単語の連接しての共起することの有意さによる

	W1	no W1
W2	a	b
no W2	c	d

- 相互情報量

$$MI = \log \frac{p(w1, w2)}{p(w1)p(w2)} = \log \frac{aN}{(a+b)(a+c)}, N = a+b+c+d$$

- χ２乗検定

$$\chi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(B+c)(b+d)}$$

- Log likelihood ratio

Contingency Matix （相互情報量と例)

◆二つの単語の連接しての共起の有意さによる

	大学	¬大学
改革	a=10	b=5
¬改革	c=5	d=980

◆相互情報量

$$MI=\log\frac{aN}{(a+b)(a+c)}$$

$$=\log\frac{10\times1000}{(10+5)\times(10+5)}$$

$$=\log 3333=8.38$$

Contingency Matix （相互情報量と例-1)

◆二つの単語の連接しての共起の有意さによる

	大学	¬大学
改革	a=10	b=100
¬改革	c=90	d=800

◆相互情報量

$$MI=\log\frac{aN}{(a+b)(a+c)}$$

$$=\log\frac{10\times1000}{(10+90)\times(10+100)}$$

$$=\log 9.09=3.18$$

相互情報量の問題点

- 二つの単語の連接しての共起の有意さによる

	大学	¬大学
改革	a=1	b=0
¬改革	c=0	d=999

- 相互情報量 $MI=\log\frac{aN}{(a+b)(a+c)} \quad =\log\frac{1\times1000}{(1)\times(1)}=9.96$
- これでは過大評価→dice係数（重み付き）

$$Dice=\log(a\times\frac{2a}{(a+b)+(a+c)})=\log(1\times\frac{2}{1+1})=0$$

$$compare\ previous \quad Dice=\log(10\times\frac{20}{15+15})=\log 6.7=2.74$$

Contingency Matix （χ2乗検定と例)

- 二つの単語の連接しての共起の有意さによる

	大学	¬大学
改革	a=10	b=5
¬改革	c=5	d=980

- χ2乗検定

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)} = \frac{1000(9800-10)^2}{15\times15\times985\times985} = 489$$

- 自由度1のχ2乗分布で棄却率は0.1%以下→有意に共起

Contingency Matix （χ2乗検定と例-1)

- 二つの単語の連接しての共起の有意さによる

	大学	¬大学
改革	a=10	b=100
¬改革	c=90	d=800

- χ2乗検定

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)} = \frac{1000(8000-9000)^2}{110\times100\times890\times900} = 0.11$$

- 自由度1のχ2乗分布で棄却率は75%以下→有意に共起

Likelihood ratio

- 仮説H1: $p(w2|w1)=p(w2|\neg w1)$
- 仮説H2: $p(w2|w1)>p(w2|\neg w1)$
- H1,H2のlikelihoodをL(H1),L(H2)とすると
- $\log\lambda=\log\frac{L(H1)}{L(H2)}$ が閾値Cより小さければ

  w1 w2は有意な連語
- *L(H1),L(H2)* の計算はちょっと面倒

計算例

$$H1: p(w2|w1) = p(w2|\neg w1) = p = \frac{a+b}{N}$$

$$H2: p(w2|w1) = p1 = \frac{a}{a+c},$$

$$p(w2|\neg w1) = p2 = \frac{b}{b+d} = \frac{b}{N-a-c}$$

$$b(k,n,x) = \binom{n}{k} x^k (1-x)^{(n-k)}$$ 二項分布

$$L(H1) = b(a, a+c, p)b(b, b+d, p)$$

$$L(H2) = b(a, a+c, p1)b(b, b+d, p2)$$

計算例

$$\boldsymbol{H1: p(w2 \mid w1) = p(w2 \mid \neg w1) = p = \frac{10+5}{1000} = 0.015}$$

$$H2: p(w2 \mid w1) = p1 = \frac{a}{a+c} = \frac{10}{10+5} = 0.67,$$

$$p(w2 \mid \neg w1) = p2 = \frac{b}{b+d} = \frac{b}{N-a-c} = \frac{5}{980+5} = 0.005$$

$$b(k,n,x) = \binom{n}{k} x^k (1-x)^{(n-k)} \quad \text{二項分布}$$

$$L(H1) = b(a, a+c, p)b(b, b+d, p) = b(10,15,0.015)b(5,985,0.015)$$

$$L(H2) = b(a, a+c, p1)b(b, b+d, p2) = b(10,15,0,67)b(5,985,0.005)$$

$$\Rightarrow \frac{L(H1)}{L(H2)} = \frac{1.39 \times 10^{-34}}{1.60 \times 10^{-18}} << 1$$

$$\Rightarrow \log\left(\frac{L(H1)}{L(H2)}\right) = -53 \Rightarrow \text{有意に共起}$$

計算例-1

$$H1: p(w2 \mid w1) = p(w2 \mid \neg w1) = p = \frac{10+90}{1000} = 0.1$$

$$H2: p(w2 \mid w1) = p1 = \frac{a}{a+c} = \frac{10}{100} = 0.1,$$

$$p(w2 \mid \neg w1) = p2 = \frac{b}{b+d} = \frac{b}{N-a-c} = \frac{100}{900} = 0.11$$

$$b(k,n,x) = \binom{n}{k} x^k (1-x)^{(n-k)} \quad \text{二項分布}$$

$$L(H1) = b(a, a+c, p)b(b, b+d, p) = b(10,100,0.1)b(90,900,0.1)$$

$$L(H2) = b(a, a+c, p1)b(b, b+d, p2) = b(10,100,0.1)b(90,900,0.11)$$

$$\Rightarrow \frac{L(H1)}{L(H2)} = \frac{6.58 \times 10^{-142}}{4.10 \times 10^{-142}} \approx 1$$

$$\Rightarrow \log\left(\frac{L(H1)}{L(H2)}\right) = 0.68 \Rightarrow \text{有意に共起ではない}$$

複合語、collocationの unithood, termhood (語彙空間)

- 複合語やcollocationの内部構造による
  - Xtract
- 語彙空間における構造
  - 共起する構造が組み合わさった場合の問題
  - 語彙の構造を反映する統計
  - C-value, 連接数

連語Collocationとは

- A sequence of two or more consecutive words
- regarded as a syntactic and semantic unit,
- Non-compositinality: its meaning cannot directly be derived grom its components
  - kick the bucket
- Non-substitutability: cannot substitute other word into its component
  - white wine ≠ yellow wine
- Non-modifiability: cannot freely modify its component
  - 奥歯にものが挟まったような ≠ 奥歯に大きなものが挟まったような

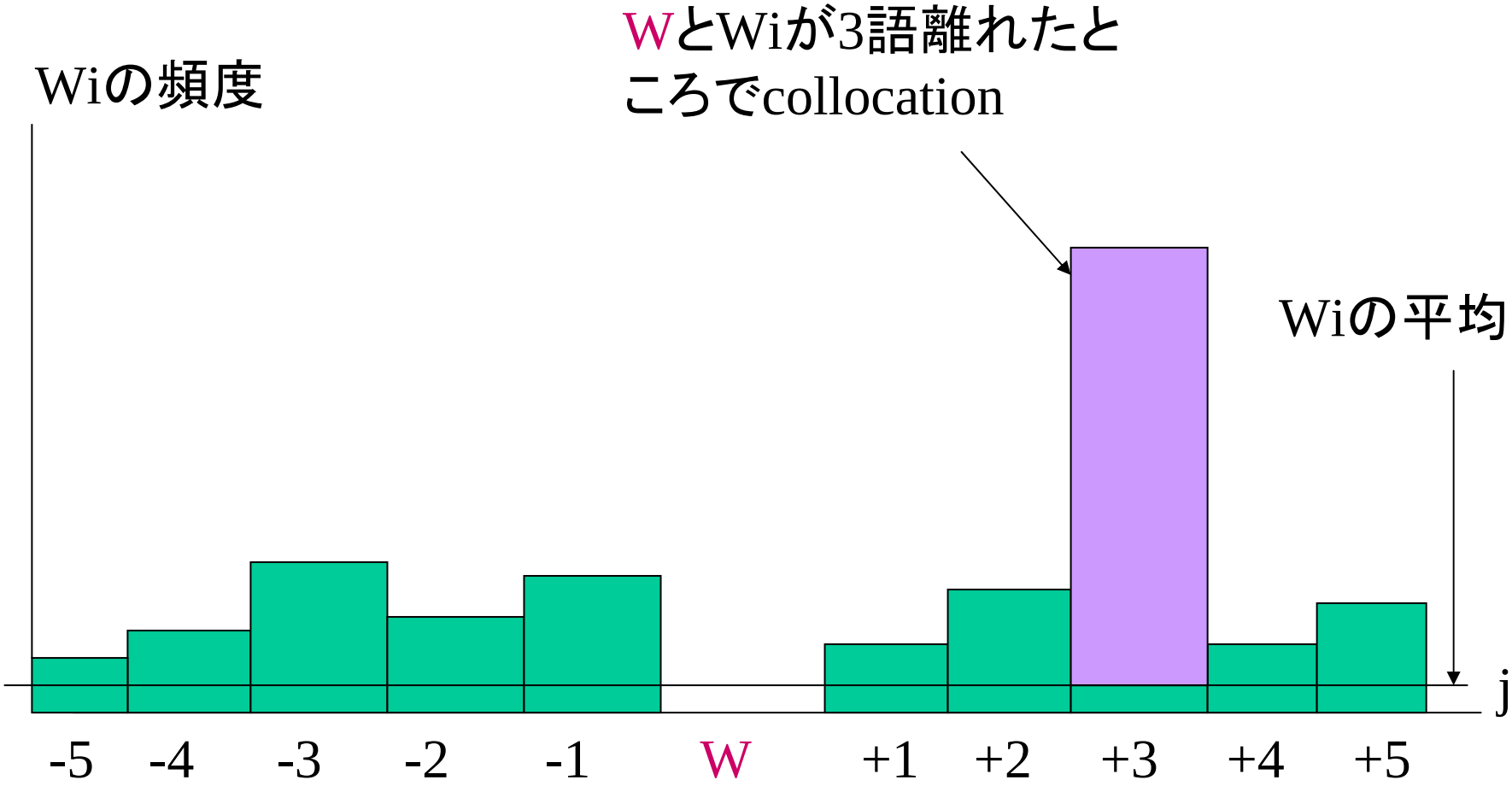
単名詞、複合名詞、連語

- 用語候補の分類＝単名詞、複合名詞、連語
- 単名詞：　これ以上分解できない名詞。専門用語のうち１０％程度
- 複合名詞：　単名詞の連接したもの。専門用語の８５％が複合名詞
- 連語：collocation：
  - その意味が要素の意味だけから構成できない
  - United states, 虎の子、　→idiom
  - 連続していなくてもよい(広義)

統計量によるCollocation抽出

- Smadja, Xtract System Computational Linguistics, 1993
- Collocation の分類
  1. 主語述語関係: make-decision, hostile-takeover
  2. 固定した名詞句: stock market, foregin exchange
  3. 句のテンプレート:The Dow Jones average fell NUMBER* points to NUBER*

XtractにおけるCollocation の捉え方



Xtract: stage1: Extracting Significant bigrams

1. Producing concordance(用語索引): タグつきコーパス+ 単語:*W* を用い、*W* を含む全ての文を抽出
2. Compile and sort: *W*と共起する単語 W_iが *W*から*j*語離れた位置に出現する頻度$freq(W_i)^j$ *(-6<j<6)*を計算
3. 統計的有意さで共起する単語対を抽出。次の3つの条件によって選択。

stage1

- *W*から-5から+5語の位置における*Wi*の頻度を *p(Wi)j (j=-5,..+5)*とする
- 条件１
  - $$k_0 < \frac{freq\ (Wi\ ) - E_{Wi}\ [\ freq\ (w_i\ )]}{\sigma_{W\ i}[Wi\ ]} = ki$$
- *k0* (予め決めた閾値) : w*i*の頻度が十分高い
- 条件２

$$Ui = (\sum_{j=-5}^{+5} (p(Wi)j - \overline{p(Wi)})^2)/10 > \theta$$

θ は閾値

  - : 近辺の頻度分布がピークを持つ。つまり固定された表現(collocation)

Stage1

- 条件３

$$p((W_i)\,j) \geq \overline{p(W_i)} + k1 \times \sqrt{Ui}$$

- 位置jにおいてW_iが有意に共起するCollocationを抽出するための条件

Stage 2: From 2-grams to N-grams

- Stage1の結果得られたbigramの周辺で m 語離れた場所(m<N)で高い確率で出現する単語を抽出して3語以上のcollocationを探す。場合によっては品詞(part of speech: 略して pos)で置き換えることにより N-gram へ拡大
- 例 composite index → The NYSE's composite index of all its listed common stockes fell *NUMBER to *NUMBER

Stage3:Adding syntax to collocations

- Stage2までですでにcollocationは得られたが、その各要素に品詞タグを付ける。
- 次に元文を構文解析し、文法役割 S, V, Oなどを与える。
- 同じ文法役割の付与がされる割合が統計的に有意に大きいなら、その役割付与をcollocationとして採用
- 構文解析しても文法役割付与ができないものはcollocationとみなさず
  - 例：〇 savings fell: SV,
  - ×savings failing: undefined

用語候補の構造と統計による方法

- unithood と termhood をより直接的に測ろうとする方法
- Ｃ－value　法 (unithood)
- 単名詞の連接における統計 (termhood)

C-value 法

- Xtractのstage2でbigramのcollocationからN-gramへ拡大した。しかし、逆方向も考えられるわけで
- Frantzi&Ananiadou96 said: ” (they try to extract) substring of other longer one(=collocations).” “ they(including Xtract) try not to extract unwanted substrings of collocations.”
- つまり、collocation の一部分もcollocation としての資格を持つなら抽出したい。
- 例：Wall Street Journal の一部の Wall Street も役立つものなら抽出したい。

C-value

- 長いcollocation:C1の一部:C2がC1と同じ頻度ならC2はcollocationとはみなさない
- a が既存のcollocationの部分でないなら、C-value(a)=(length(a)-1)n(a),　ただしn(a)はaの頻度
- aが既存のcollocationの部分なら C-value(a)=(length(a)-1)(n(a)-t(a)/c(a))
- t(a)はaが長いcollocation 内部に現れる頻度、c(a)は長いcollocationの異なり数

C-valueの計算　作例

- 例:コーパスから次の出現回数が分かったとする。
- 単語　トライ　グラム(3回)、　トライ　グラム　統計(2回)
  クラス　トライ　グラム(1回)、トライ　グラム　獲得(1回)
  文字　トライ　グラム(1回)、トライ　グラム(４回)
- ここで「トライ　グラム」の　C-value を計算する。
- length(トライ　グラム)=2　　n(トライ　グラム)＝１２回
- t(トライ　グラム)＝8回　　c(トライ　グラム)＝5種
- C-value(トライ　グラム)
  ＝(length(トライ　グラム)-1)(n(トライ　グラム)-t(トラ..)/c(トラ..))
  ＝(2-1)(12-8/5)=10.4
- C-value(単語トライグラム)＝(3-1)3=6

C-valueの抽出実例

- Examples:
  - WALL STREET JOURNAL,
  - Staff Reporter of The Wall Street Journal,
  - Wall Street,
  - of its, it is, because of
- C-value は length(a) に比例するので、長いcollocationが優先される傾向がある。

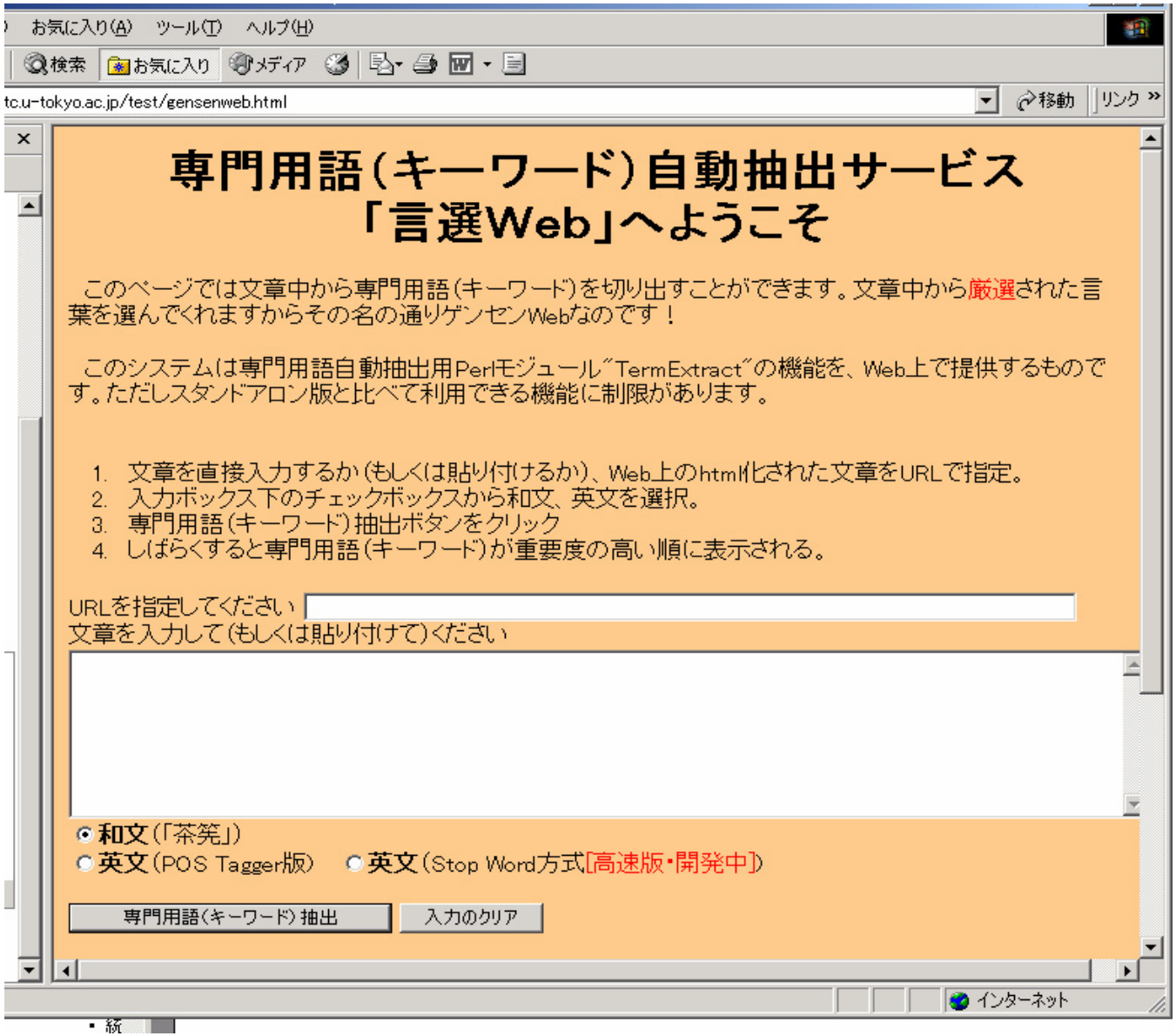
言選Web
Webからの専門用語抽出

- 小さなテキストからテキストを特徴付ける専門用語を抽出
- 多言語に適用可能

  →対訳の候補を求められる
- 順位が付いている

単名詞の連接による方法

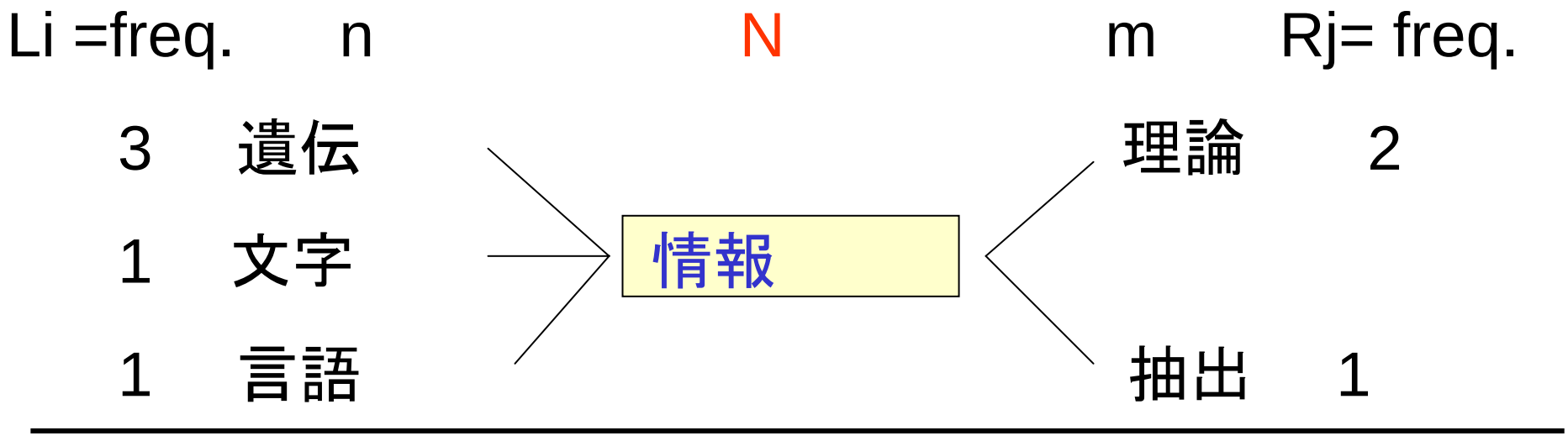
- C-valueが長いcollocationからその部分を取り出す方法であったのに対して、単名詞が複合語を作る場合の連接数により、まず単名詞の重要度を求める方法。bigramから始めるXtractとも異なる。
- 単名詞に重みを与え、それを組み合わせる方法はユニーク（見返りは、一度に抽出できるのが同一分野の用語に限定される点）
- 文書集合における頻度ではなく、語彙集合における複合語の構成に関する情報を利用。Webの1ページくらいの小さなテキストでもそこそこ機能する。


お気に入り(A) ツール(T) ヘルプ(H)
検索 お気に入り メディア
tc.u-tokyo.ac.jp/test/gensenweb.html
移動 リンク
専門用語(キーワード)自動抽出サービス
「言選Web」へようこそ
このページでは文章中から専門用語(キーワード)を切り出すことができます。文章中から厳選された言葉を選んでくれますからその名の通りゲンセンWebなのです！
このシステムは専門用語自動抽出用Perlモジュール"TermExtract"の機能を、Web上で提供するものです。ただしスタンドアロン版と比べて利用できる機能に制限があります。
1. 文章を直接入力するか(もしくは貼り付けるか)、Web上のhtml化された文章をURLで指定。
2. 入力ボックス下のチェックボックスから和文、英文を選択。
3. 専門用語(キーワード)抽出ボタンをクリック
4. しばらくすると専門用語(キーワード)が重要度の高い順に表示される。
URLを指定してください
文章を入力して(もしくは貼り付けて)ください
和文(「茶筅」)
英文(POS Tagger版)
英文(Stop Word方式[高速版・開発中])
専門用語(キーワード)抽出
入力のクリア
インターネット

言選Webのアルゴリズム

- ◆多くの複合名詞の要素になっている基本名詞（単名詞）ほど重要度が高い
  - ◆多くの複合名詞（＝多くの専門分野概念）を形成する要素になる単名詞（＝基礎概念）ほど重要　というアイデア
- ◆テキストを離れて語彙だけが形成する空間で重要度を計算するのでテキストの大きさに依存しにくい。

単名詞のスコア付け



LN(情報)=5　n=3　　m=2　　RN(情報)=3

遺伝情報、遺伝情報抽出、文字情報、言語情報理論、

遺伝情報、情報理論

複合語のスコア付け
相乗平均法

$$LR(CN) = \left\{ \prod_{i=1}^{L} \left[\left(LN(N_i) + 1 \right) \cdot \left(RN(N_i) + 1 \right) \right] \right\}^{\frac{1}{2L}}$$

$$CN = N_1 N_2 \ldots\ldots N_L$$

GM(CN)は相乗平均なので、複合語CNの長さ
（＝要素となる基本単語数）に依存しない重要度となる

出現頻度も考慮したスコア: FGM(CN)

if *CN* が独立に出現

then $FLR(CN) = f(CN) \times LR(CN)$

where *f(CN)* は*CN*の独立出現頻度

（= CN がより長い複合語の一部とはならずに出現した頻度）

Ex. LR（情報）$=((5+1)\text{x}(3+1))^{1/2}=4.9$

if f(情報)= 5

FLR(情報)=24.5

言選Webを使えば、小さなテキストから重要語が求まる。

そこで、

- 単独のWEBページからでもキーワードが求まる。

中国語への応用

- 言選Webの適用は単語か文字か？

「言选Ｗｅｂ」（中文·停止语方式版）

本网页能从文章中抽出专业用语（关键字）。

本系统提供在线专业用语自动抽出Perl模块"TermExtract"的功能。其与单独运作(stand-alone)版相比，虽然利用功能有所限制，但有检索方便的

1. 直接输入（或剪贴）文章，或以URL指定网上的html文档。
2. 选择中文（GB）或中文（UTF-8）。
3. 点击专业用语（关键字）抽出按钮。
4. 专业用语（关键字）将以重要度的顺序抽出。

请输入URL

请输入（或剪贴）文章

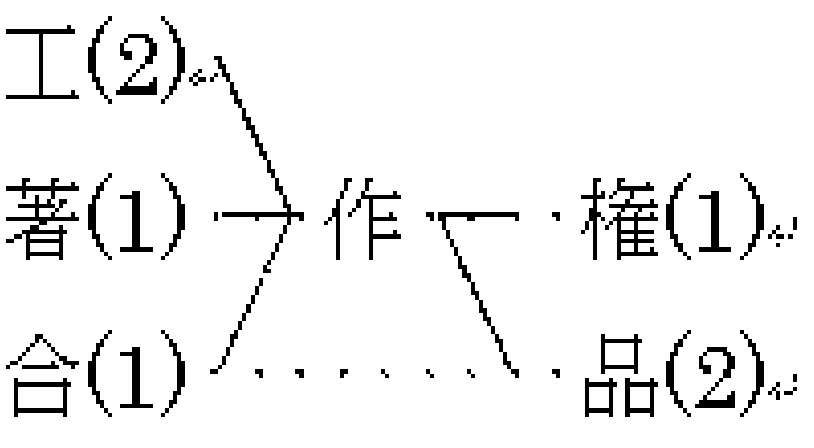
◉ **中文**（GB） ○ **中文**（UTF-8）

专业用语（关键字）抽出 | 清除 | ICTCLAS版

Key word only

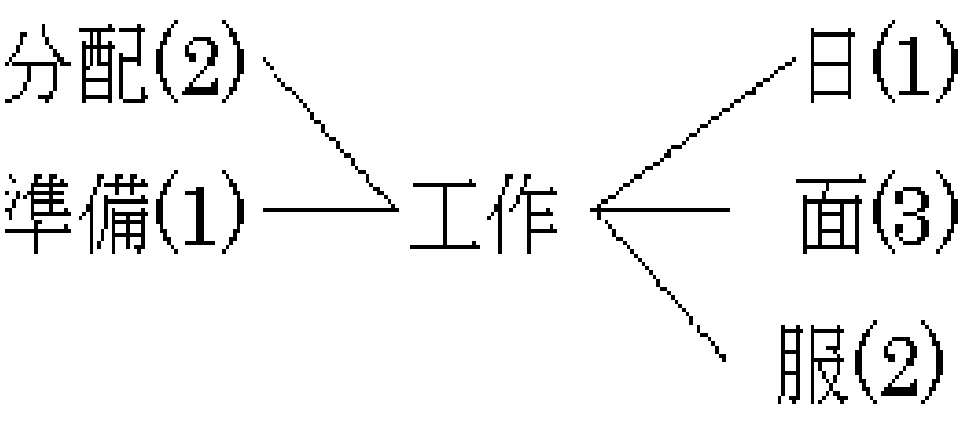
文字ベースのFLRの例

工作、著作権、合作、工作、作品、作品



LN(作）＝４　　RN(作）＝３

単語ベースのLRの例



- LN(工作）＝3、RN(工作）＝6
- 形態素解析による単語の切り出しが必要

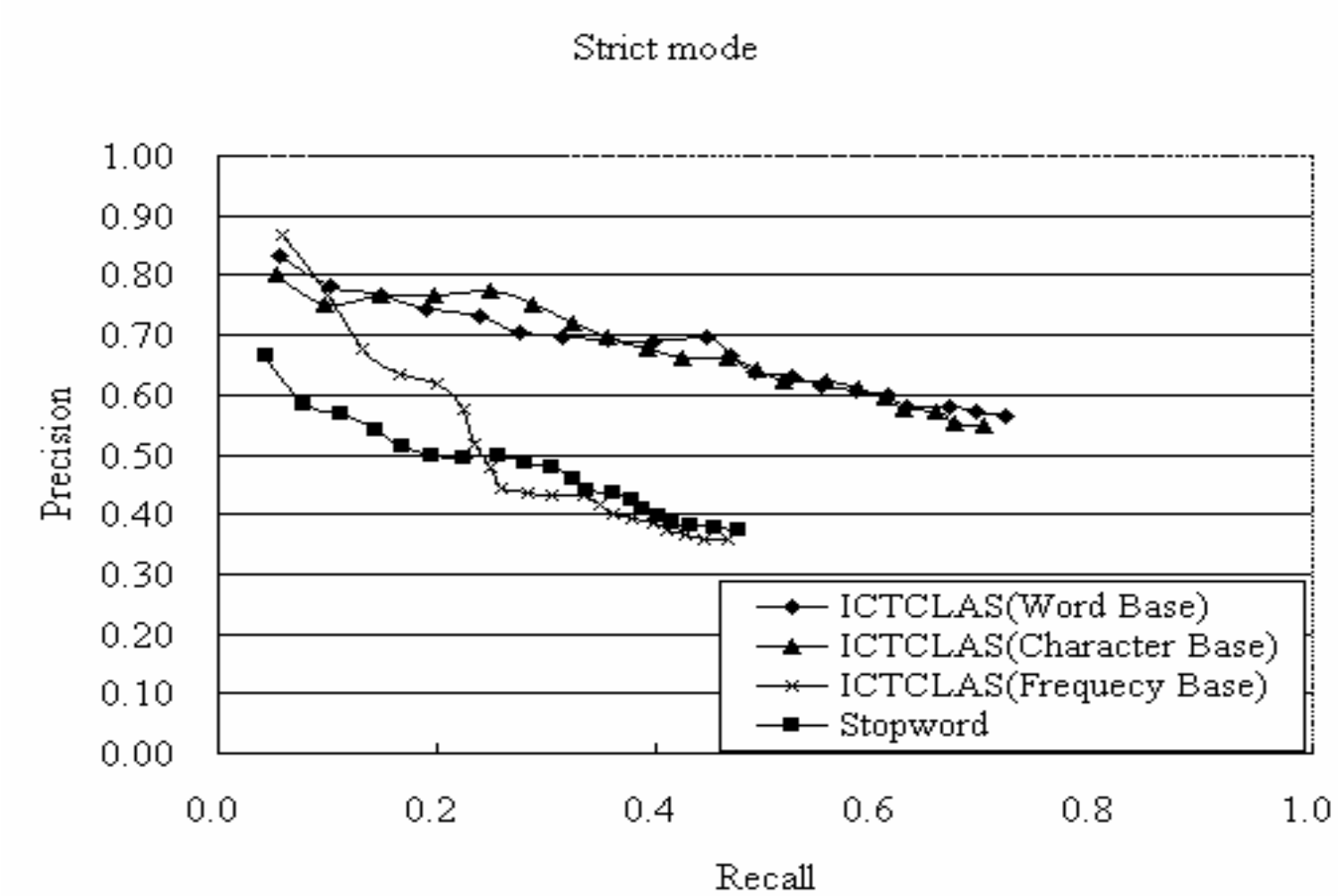
ICTCLASによる形態素解析と用語候補の抽出ルール

- ✓MWU <-- [ag a]* [ng n nr ns nt nz nx vn an i j]+
- ✓MWU <-- MWU?b [ng n nr ns nt nz nx vn an i j]+
- ✓MWU ← [ag a]+ [u k] MWU
- ✓MWU ← MWU (u|k|*he-2*|*yu-3*) MWU

評価した方法

- (1) Stop-word による候補切り出し + 文字ベースの FLR.
- (2) POS tag による候補切り出し＋文字ベースの FLR
- (3) POS tag による候補切り出し＋単語ベースの FLR
- (4) 単純な頻度による方法

人民日報30記事で評価



Bootstrap法

- Demetriou00(LREC2000)
  1. Initialisation
     1. Extract initial set of contextual patterns (left and right n-grams n=6) using seeds
     2. Identify significant patterns(χ^2検定　0.5%）
     3. Terminate if no significant patterns exist
  2. Name extraction
     1. Apply significant patterns and extract new names
     2. Terminate if no new names are found
  3. Pattern generation
     1. Match the new names and extract new patterns
     2. Identify significant patterns
     3. Terminate if no significant patterns exist
  4. Go to step 2

Bootstrap法 つづき

- 医学文献からの蛋白質名前抽出
- パターン例：of human, 3-dimensinal structure of , the bacterial,….
- 結果
  - Original data:Recall=31%, precision=97%
  - Original+new terms: R=38%,P=96%
  - Original+newt terms+grammar rules: R=90%,P=96%
- 知見
  - 専門分野文献のほうがBNCよりperplexity低い
  - 専門分野文献のほうが文法的には変化が少ない
- NEタスクに近い方法

２言語コーパスを利用する方法

- Beatrice Daille 94
- 英語（フランス語）のコーパスから統計的手法で単名詞のペアを抽出し、
- ペアからいろいろな compound を生成し、
- Compound 生成の例：(interference, level) → interference level(s), level of interference(s)
- いろいろな統計量(Log-likelihood, MI など）を組み合わせてみたが、失敗

２言語コーパスを利用する方法

- Beatrice Daille 94
- 英語（フランス語）のコーパスから統計的手法で単名詞のペアを抽出し、
- ペアからいろいろな compound を生成し、
- Compound 生成の例：(interference, level) → interference level(s), level of interference(s)
- いろいろな統計量(Log-likelihood, MI など）を組み合わせてみたが、失敗

Dalli の方法の続き

- そこで　aligned な英仏コーパスによって、一方の言語のコーパスから生成したターム候補が正しいなら相手側の aligned sentence にも等価なタームが頻繁に存在するという仮説によって正しいタームを選ぶ。
- Top 500 → 80%, Top 1000 → 70% Precision
- ターム候補を作ってから二言語コーパスを利用するというのは新しい考え。中川も似た方法で日英対訳をNTCIR1,2 corpus から作った。

用語を拡大する

- 抽出した用語だけでは、不十分なこともある。
  - 例えば、情報検索で使うキーワードは、利用者が入力したキーワードを拡大して使うとよいこともある。
  - 1語の場合：ソート → ソートアルゴリズム、配列ソート
  - 2語の場合：日本語、解析→日本語構文解析、日本語語彙解析、日本語形態素解析

用語を拡大する

- 1語の場合：ソート → ソートアルゴリズム、配列ソート
- 2語の場合：日本語、解析→日本語構文解析、日本語語彙解析、日本語形態素解析

- どのような単語を使って拡大するか？
  - シソーラスを使って関連語句を使って拡大
  - 文法的に意味ある拡大
    - 配列ソート →「配列　を　ソートする」という文を短縮した複合語

構造的に構成する方法論

- derivational morphology
  - 既存の用語の構成要素（単名詞）のvariationを作り、未知の用語を生成
  - 既存の用語の文法的に正しい結合の規則により生成
    - フランス語の例：

      N1 de N3 + N2 de N3 → N1 et N2 de N3

      Assemblage et deassemblage de paquet

FASTER

- Jacquemin&Rayoute94 (SIGIR94)
- 文脈自由文法で複合語の用語を生成する規則を記述
- Postive meta-ruleで規則を拡大ないし洗練
  - Coordination: (X1→X2 X3 X4) = X1 → X2 C5 X6 X3 X4
    - Inflammatory ***and erosive*** joint disease
  - Insertion:(X1→X2 X3 X4) = X1 → X2 X5 X3 X4
    - Impaired intravenous glucose tolerance
  - Permutation: :(X1→X2 X3 X4) = X1 → X4 X5 X6 X7 X2 X3
    - Disease of the central nervous system [Nervous system disease]
- Negative meta-rule で生成された（良くない）候補を排除
  - Coordination :

    (X1 → X2 X3) = X1 → X2 C4 X5 X3: <X2 number> = plural

    （評価実験で5%位偽用語を排除した）
  - × cells or fatal cultures ← Cell cultures
- 9MBの医学コーパスから31,428用語、そしてFASTERによってさらに8,747語の正しいvariationsを抽出。内訳はPermutation 48%, insertion 43%, coordination 9%

言語学的構造から Collocation’s variants生成

- Jacquemin　SIGIR　94,97,99,
- 与えられた複数の単名詞から
  - 単名詞の形態論的、および意味論的変化形
  - 両者を含む統語構造
- を使って、それらの単名詞を含む variation を生成する。（inflection rich な西欧の言語に即したヨーロッパ的なやり方）
  - FASTER　というシステムとして公開

Derivational Morphology

- Jacquemin97(SIGIR)
  1. 単語後部を切り取って一致する部分を求める。例 immuniz-(ation,ed)
  2. Two-words term の各々が1.で求め一致部分から派生する例を作る。ただし、ここでsuffixとしては後min３文字、複合語に含まれる単語数は2としている
     例　continue（実験による最適値）
     - 例　active immunization , actively immunized
  3. 複数の複合語(class)から共通のsuffix(signatureと呼ぶ)を取り出す。
     - 例　(continuous measure-ment) (continuous-ly measure-d)→ {(ε ,ment),(ly,d)} そしてこれを生成に使う
     - 例えば、{(ε , ing),(ly,ed)}により
       diffuse scattering → diffusely scattered

Derivational Morphology

3. 続き : filter out
   - F = (class内の単名詞の語幹文字数の平均値)/(signatureの最大値)
   - F>1のclassのみ残す。つまり、変化語尾(＝signature)が相対的に長いclassは捨てる
4. Classのクラスタ化
   - Class間の距離の近いものをまとめる。Signatureの最後尾の文字ほど一致する場合の重みを大きくするような距離の定義による

Collocation's variants生成 Jacquemin 99(ACL99)

- Morphological family's example
  - F_M(measurement)={commensurable,countermeasure, tape-measure,measure, ….}
- Semantic family's example
  - WordNet: F_{SC}(speed)={speed,speeding, hurrying,velocity,amphetamine,..}
  - Word97: F_{SL}(speed)={speed, rapidity, celerity,..}
- 生成規則の例:
  - N1Prep N2 → F_M(N1)Adv? A? Prep Art? A? F_{SC}(N2)
  - 例: composition du fruit → compse'chimiques de la graine (chemical compound of the seed)
  - この他に Coordination, Modification, Permutation, VP化、NP化 の規則あり。

Collocation’s variants生成

- 例
  - Pressure decline → pressure rise and fall
  - Angular measurement → angles measure
- 形態素と統語規則だけだと80%近い精度
- 意味論的規則を混ぜると50%以下（しかし、この方法での生成variantsは数%以下の極少数）
- この方法でテキストに現れたcollocationの3倍以上の量のvariantsを生成

相互情報量などによる方法－１

- Su-Wu-Chan (ACL94)
- 単語2-，３-gramを複合語候補とする。
- 選択の基準は相互情報量MI、相対頻度（RFC)、品詞パターンL_i
- MIはbi-gram $I(x;y)=\log_2 \frac{P(x,y)}{P(x)\times P(y)}$
- 3-gram $I(x;y;z)=\log_2 \frac{P(x,y,z)}{P(x)P(y)p(z)+P(x)P(y,z)+P(x,y)P(z)}$

相互情報量などによる方法－２

- 品詞パターンは L_i=[n,n]など
- 文脈までいれると　L'_{ij}=[adj (n n) n]など
- 以上３種類の情報を総合すると、$M_{c\ (nc)}$がn-gramが（非）collocationから生成されたという事象とすると

$$P(\mathbf{x}|M_c)\times P(M_c)$$
$$\approx \prod_{i=1}^{n}[P(MI_{Li}, RFC_{Li}|M_c)\prod_{j=1}^{ni}P(L'_{ij}|M_c)]\times P(M_c)$$

- MI,RFCの項は、正規分布を仮定すれば、テストセットから平均、分散を求めれば推定可能

相互情報量などによる方法－３

- 品詞パターンの項は、bigramの場合

$$P(L'_{ij}|M_c) = P(C_0, C_1, C_2, C_3|M_c)$$
$$\approx P(C_3|C_{2,}M_c) \times P(C_2|C_{1,}M_c) \times P(C_1|C_{0,}M_c) \times P(C_0|M_c)$$

- 3-gramの場合もほぼ同様
- 確率はテストセットデータから求め、unseenデータの判断は、likelihoodによる。ただし、*c*は複合語の場合、*nc*は複合語でない場合

$$\lambda = \frac{P(\mathbf{x}|M_C) \times P(M_C)}{P(\mathbf{x}|M_{nc}) \times P(M_{nc})}$$

相互情報量などによる方法－４

- Suらの実験では人手で修正された形態素解析済みのコーパスを使う。
- bigramで　recall=0.977,prec=0.445 (training set)
  recall=0.962, prec=0.482(test set)
- 3-gramで

  recall=0.976,prec=0.402 (training set) ,
  recall=0.966, prec=0.396(test set)
- 抽出例：dialog box, mail label, main document, datafile, file menu, World User’s guide, Microsoft Word User’s, Template option button, new document base, File name box

C-valueの拡張

- NC-value (Frantzi et al 2000, Maynard et al 2001)
  - 対象にしているターム*a* の C-value と *CF(a)*を線形結合 C-value*0.8＋ *CF(a) *0.2*

$$CF(a) = \sum_{w \in Ca} F(a,w) F_w / n_w$$

  - *Ca* は*a* の文脈に現れるタームの集合
  - *F(a,w)* は*a* の文脈に現れた*w*の出現回数
  - *Fw*は*w*と同じ文脈に現れたターム数
  - n_w は*w*のコーパス中での総出現回数