情 報 学 基 礎 61−16
自 然 言 語 処 理 142−16
（2001. 3. 6）

テキストコーパスにおける特徴語抽出のための分析ツール

相澤 彰子

国立情報学研究所

(akiko@nii.ac.jp)

本稿では，形態素情報を手がかりに取り出した複合名詞を順序付き単語リストとみなして共起関係を分析する方法について考察する．本稿で語の特徴度の尺度として用いるのはFQ値と呼ぶ量で，対数尤度比に類似の尺度である．本稿では，任意長の単語列として定義される各語のFQ値と頻度に基づき，結合度，出現度，前接度，後接度，文脈度，重要度と呼ぶ各尺度を定義し，これら異なる観点に基づく語のランキングを効率的に行うための計算法，および上限と下限値の理論的な解析を行う．また実際のテキストコーパスを用いた計算例を示す．

A Method for Analyzing Feature Terms of Text Corpora

Akiko AIZAWA

National Institute of Informatics

This paper discusses a method for analyzing co-occurences of terms using sequential word lists automatically generated using the POS tags output of an existing morphological analyzer. In this paper, FQ (feature quantity)-value, that is similar to the log likelihood ratio, is first introduced as a basic metric of a term, and then, different measures such as connectivity, appearance, precedence, succession, contextuality, importance are defined. Next, a method for effciently calculating the values of these measures is shown, together with the equations to give the theoretical upper and lower bounds. An example of the calculation is also shown using an actual text corpus.

1 はじめに

本稿では，与えられたテキストコーパスから特徴的な語を抽出するための数量的な尺度の定義および効率的な計算法について，現在の検討結果を報告する．まず本稿で想定している処理全体の流れを図1に示す．始めに形態素情報を手がかりに，最長一致により複合名詞をセグメントとして切り出す．次に，その構成要素である（部分）単語列の共起について数値的な分析を行う．最後に分析結果に基づき，すべての単語列を指定された尺度にしたがってランキングする．

簡単な例を以下に示す．たとえば，

「テキストコーパスにおける特徴語抽出のための分析ツール」

というテキストは形態素解析ツールChasen Ver.2.02[1]により

テキスト／名詞−一般，コーパス／名詞−一般，における／助詞−格助詞−連語，特徴／名詞−一般，語／名詞−接尾−一般，抽出／名詞−サ変接続，の／助詞−連体化，ため／名詞−非自立−副詞可能，の／助詞−連体化，分析／名詞−サ変接続，ツール／名詞−一般

のように単語に分割される．これを予め定義した品詞パターンと照合することにより，

S_1 （テキスト， コーパス）
S_2 （特徴， 語， 抽出）
S_3 （分析， ツール）

のような順序つき単語リストを構成して，

テキスト，コーパス，テキストコーパス，特徴，特徴語，抽出，特徴語抽出，分析，ツール，分析ツール

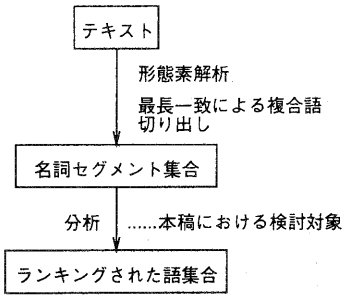

図 1: 想定する処理の流れ

等の語を数量的に評価し，定義した尺度に基づきランキングする．

以下，形態素解析ツールの出力として得られる語単位を「単語」として，各単語を w，与えられたテキスト集合中に出現する全ての単語の集合を $\mathcal{W}$ で表記する．また分析の対象とする「語」は，複合名詞に相当する単語列であり，任意の k（$k \geq 1$）個の単語から構成される順序付きリストで表されるものとする．すなわち語を T として，

$$T = (w_1, \cdots, w_k) \qquad (w_i \in \mathcal{W})$$

である．語 T を構成する単語の数を以下便宜的に T の「長さ」と呼ぶ．また，$w_1, \cdots, w_k$ を略して w_1^k のように表記する．さらに，テキストから取り出した最長一致による複合名詞を「(名詞) セグメント」と呼び，S で表記する．

本稿で想定している語の抽出手順は，上記のように任意長の n グラムに基づく単純なものである．ただし例からも容易にわかるように，最長一致により抽出したセグメントからは，重複する多数の部分単語列を語として取り出すことが可能であり，これらの語すべてが分析の対象となる．本稿で特に検討を試みるのは，(1) これら長さの異なる語を数量的に比較するための数量尺度を異なる複数の観点に基づき定義し，相互の関係を明確にすること，(2) 大量のテキストを分析するための効率的な計算法を検討すること，の 2 点である．

以下，**2.** では共起関係の抽出で用いられる 2×2 分割表に対する統計尺度について概観したのち，FQ 値と呼ぶ特徴量を定義する．**3.** では，各語の出現頻度と FQ 値を用いて，結合度，出現度，前接度，後接度，文脈度，重要度という異なる観点に基づく語の数量尺度を定義する．**4.** では，これら異なる尺度の計算を効率に行うための木構造表現について述べ，FQ 値に関する上限値および下限値の計算式を示す．最後に，**5.** で実際のテキストコーパスを用いて計算を行った例を示し，今後の課題について述べる．

2 分割表と FQ 特徴量

2.1 2×2 分割表における各種の統計尺度

従来より特徴語抽出のために用いられてきた尺度の多くは，図 2 に示す 2×2 分割表に基づき，2 つの属性間の相関（従属性）を評価するものである．具体的には，語 u が「生じる」「生じない」（図では $i = 1, 2$ で表記），語 w が「生じる」「生じない」（$j = 1, 2$ で表記）の組合わせによる 4 事象について，頻度 f_{ij} $(i, j = 1, 2)$ あるいは確率 p_{ij} $(i, j = 1, 2)$ に基づき，両者の共起の度合を評価する．

	w j=1	$\overline{w}$ j=2	$\sum$
u i=1	f_{11}	f_{12}	$f_{1\cdot}$
$\overline{u}$ i=2	f_{21}	f_{22}	$f_{2\cdot}$
$\sum$	$f_{\cdot 1}$	$f_{\cdot 2}$	F

(a) 2×2 分割表

	w j=1	$\overline{w}$ j=2	$\sum$
u i=1	p_{11}	p_{12}	$p_{1\cdot}$
$\overline{u}$ i=2	p_{21}	p_{22}	$p_{2\cdot}$
$\sum$	$p_{\cdot 1}$	$p_{\cdot 2}$	1

(b) 確率による表現

図 2: 2「語」を 2 属性に対応させた 2×2 分割表

ここで，代表的な統計尺度を分類すると以下のようになる．

(1) 2 語が互いに共起する頻度（f_{11}）あるいは確率（p_{11}）が大きいほど共起の度合が強いとみなす方法．

(2) ダイス係数（$\frac{2f_{11}}{f_{1\cdot}f_{\cdot 1}}$）や特殊相互情報量（$log\frac{p_{11}}{p_{1\cdot}p_{\cdot 1}}$）のように，2 語が共起する事象の出現の度合を，独立性を仮定する場合と比較することにより評価を行う方法．

(3) χ^2 統計量（$\sum_i\sum_j(f_{ij} - \frac{f_{i\cdot}f_{\cdot j}}{F})^2/\frac{f_{i\cdot}f_{\cdot j}}{F}$），対数尤度比 χ^2（$2\sum_i\sum_j f_{ij}log\frac{Nf_{ij}}{f_{i\cdot}f_{\cdot j}}$），相互情報量（$\sum_i\sum_j p_{ij}log\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}}$）のように，4 事象すべてを考慮して，独立性を仮定する場合との違いを評価する方法．

(4) 修正ダイス係数（$(logf_{11})\frac{2f_{11}}{f_{1\cdot}f_{\cdot 1}}$），本稿で用いる FQ 値（$p_{11}log\frac{p_{11}}{p_{1\cdot}p_{\cdot 1}}$）のように，(2) の共起の度合に出現頻度や確率による重みをかけあわせる方法．

従来より，(1) を用いると高頻度語が過剰に評価され，(2) を用いると低頻度語の重みが強くなりすぎることが知られている．これより (3) または (4)（あるいは Odds Ratio，ϕ^2，t スコア等，ここでの分類に明確に対応しないが類似の尺度）が用いられることになるが，通常の場合は $f_{11} \ll f_{12}, f_{21}, f_{22}$ であることから，いずれの尺度を用いても評価値の，頻度や出現確率に対する傾向は同様であり，本質的に大きな相違はないものと考えられる．

ここで，2×2 分割表に基づく評価はあくまで 2 語間の共起の度合を調べるためのものであり，任意の k 単語の共起を評価する場合には，$(k-1)$ 語と残りの 1 語の共起関係というように，2 語間の関係に置換えて評価を行うことが必要である．このような操作によらない評価尺度として，C-value [2] や Nested Collocation [3] などの経験的な尺度など，前後に接続する語の分布（文脈）を考慮した尺度がある．ただし，これらの尺度は，2×2 分割表に基づく統計尺度とは異なる観点に基づき，独立に定義されるものである．

2.2 FQ 値の定義式

本稿では，上記で (4) のグループに属する FQ 値の定義を任意の k 語の共起に拡張することで，語を構成する単語の共起と語の文脈を同時に評価する方法について検討する．「FQ 値」(Feature Quantity Value) [4][5] は確率に特殊相互情報量をかけあわせた尺度で，2 語間の共起について考える場合には，

$$\mathcal{F}(u,w) = P(u,w) log \frac{P(u,w)}{P(u)P(w)} \tag{1}$$

のように定義され，$\sum_u \sum_w \mathcal{F}(u,w)$ が 2 語間の（一般）相互情報量に等しくなるように構成されている．FQ 値の計算では，$(u,\overline{w})$，$(\overline{u},w)$，$(\overline{u},\overline{w})$ といった事象を考慮しないことから，n 語間の共起関係の分析に容易に拡張することが可能である．具体的には語 w_1^k に対する FQ 値を次式で定義する．

$$\mathcal{F}(w_1^k) = P(w_1^k) log \frac{P(w_1^k)}{P(w_1) \cdots P(w_k)} \tag{2}$$

ただし $k=1$ のとき，すなわち語が 1 個の単語であるとき $\mathcal{F}(w_1) = 0$ とする．

出現確率 $P(w_1^k)$ は，単語列 w_1^k の出現回数 $freq(w_1^k)$ と総のべ単語数 $F = \sum_{w_i \in \mathcal{W}} freq(w_i)$ から 以下で定める．

$$P(w_1^k) = \frac{freq(w_1^k)}{F} \tag{3}$$

またセグメント内で，単語 $w_0 \in \mathcal{W}$ が w_1^k に前接することを，特に明示的に (w_0, w_1^k) と表記する．同様にセグメント内で，単語 $w_{k+1} \in \mathcal{W}$ が w_1^k に後接することを，特に明示的に (w_1^k, w_{k+1}) と表記する．w_1^k がセグメントの先頭で出現する確率を $P(\emptyset, w_1^k)$ で表記することにすると，

$$P(w_1^k) = P(\emptyset, w_1^k) + \sum_{w_0 \in \mathcal{W}} P(w_0, w_1^k) \tag{4}$$

が成り立つ．同様に，w_1^k がセグメントの末尾で出現する確率を $P(w_1^k, \emptyset)$ で表記することにすると，

$$P(w_1^k) = P(w_1^k, \emptyset) + \sum_{w_{k+1} \in \mathcal{W}} P(w_1^k, w_{k+1}) \tag{5}$$

が成り立つ．定義より明らかに $P(w_0, w_1^k) \leq P(w_1^k)$ かつ $P(w_1^k, w_{k+1}) \leq P(w_1^k)$ となる．具体的には，語「情報」の出現確率は，「情報・システム」「情報・検索」等「情報」の後ろに任意の 1 語を加えて構成される各語の確率の総和より大きく，その差分は「情報」という単語で終了する語の出現確率に対応する等である．以下，w_1^k にさらに語が前接，後接する比率を前接・後接比率と呼び $\delta_p(w_1^k)$ および $\delta_s(w_1^k)$ で表記する．すなわち，

$$\delta_p(w_1^k) = \frac{P(w_1^k) - P(\emptyset, w_1^k)}{P(w_1^k)} = \frac{\sum_{w_0 \in \mathcal{W}} P(w_0, w_1^k)}{P(w_1^k)} = \frac{\sum_{w_0 \in \mathcal{W}} freq(w_0, w_1^k)}{freq(w_1^k)} \tag{6}$$

および

$$\delta_s(w_1^k) = \frac{P(w_1^k) - P(w_1^k, \emptyset)}{P(w_1^k)} = \frac{\sum_{w_{k+1} \in \mathcal{W}} P(w_1^k, w_{k+1})}{P(w_1^k)} = \frac{\sum_{w_{k+1} \in \mathcal{W}} freq(w_1^k, w_{k+1})}{freq(w_1^k)} \tag{7}$$

である．定義より明らかに，$0 \leq \delta_p(w_1^k) \leq 1$，$0 \leq \delta_s(w_1^k) \leq 1$ となる．

3 語を特徴づける異なる数値尺度の定義

3.1 結合度

式 (2) の FQ 値は，w_1^k が互いに共起する場合の情報量と，それぞれ独立に生起する場合の情報量の和との差分を評価するものであり，語を構成する単語の共起の度合いを示す「結合度」(Con) の尺度に対応している．すなわち，

$$Con(w_1^k) = \mathcal{F}(w_1^k) = P(w_1^k) log \frac{P(w_1^k)}{P(w_1)\cdots P(w_k)} \quad (8)$$

と定義する．この値が大きいほど，テキスト中で w_1^k のパターンが特徴的に出現していると考えられる．式 (8) において，もし出現確率 $P(w_1^k)$ が同じであるならば，対数分母の $P(w_1)\times\cdots\times P(w_k)$ の値が小さいほど（直観的には長い語ほど），その値は大きくなる．また対数分母の値が同じであれば，$P(w_1^k)$ の値が大きいほど（頻度が高い語ほど），結合度の値は大きくなる．前述のように，語が 1 つの単語から構成される場合は，結合度はゼロと定義される．

3.2 出現度

情報量の観点から評価した語の手がかりとしての有用性を「出現度」(App) として，語の出現確率 $P(w_1^k)$ に基づき次式で定める．

$$App(w_1^k) = -P(w_1^k) log P(w_1^k) \quad (9)$$

ここで $-logP(w_1^k)$ は w_1^k の生起による情報量である．式 (9) の値は $P(w_1^k)$ に関して $P(w_1^k) \leq 1/e(\sim 0.38)$ の範囲で単調増加となる．通常に想定されるテキストでは単一の語の出現確率が $1/e$ 以上になることはないと考えられることから，一般に頻度が高い語ほど式 (9) の値は大きくなるといえる．すなわち，出現度による順位は頻度による順位と同じとみなしてよい．

3.3 前接度

ある語の直前に別の語を加えて新たな語を構成するという観点から，「前接度」(Pre) の尺度を以下で定義する．

$$Pre(w_1^k) = \sum_{w_0 \in W} Con(w_0, w_1^k) = \sum_{w_0 \in W} \mathcal{F}(w_0, w_1^k) \quad (10)$$

すなわち，語 w_1^k に前接して新たに単語を 1 つ加えて得られるすべての語について，結合度の総和をとった値である．ここで式 (2) の定義により式 (10) を書き改めると以下の通りとなる．

$$\begin{aligned} Pre(w_1^k) &= \sum_{w_0 \in W} \left[P(w_0, w_1^k)\, log \frac{P(w_0, w_1^k)}{P(w_0)P(w_1)\cdots P(w_k)} \right] \\ &= P(w_1^k) \sum_{w_0 \in W} \left[P(w_0|w_1^k) \times log \frac{P(w_0|w_1^k)}{P(w_0)} \right] \\ &\quad + \left[\sum_{w_0 \in W} P(w_0, w_1^k) \right] \times log \frac{P(w_1^k)}{P(w_1)\cdots P(w_k)} \\ &= P(w_1^k) \mathcal{D}\big(P(W_0|w_1^k) || P(W_0)\big) + \delta_p(w_1^k) \mathcal{F}(w_1^k) \end{aligned} \quad (11)$$

ただし $\mathcal{D}(\cdot||\cdot)$ はカルバックライプラー情報量であり，便宜的に「語がセグメントの先頭である場合，その前には任意の語が来てよい」とみなし，$P(\emptyset|w_1^k) = P(\emptyset)$ とした．また 2 行目では $P(w_0, w_1^k) = P(w_0|w_1^k)P(w_1^k)$ を，3 行目では式 (6) および式 (8) を用いた．

式 (11) の第一項は，2 つの確率分布 $P(W_0|w_1^k)$ と $P(W_0)$ のカルバックライプラー情報量による距離に w_1^k の生起確率を乗じた値であり，w_1^k が与えられた場合に前接語がどの程度制約されるかという評価に対応している．また第二項はすでに定義した通り w_1^k の結合度に前接比率 $0 \leq \delta_p(w_1^k) \leq 1$ を乗じた値であり，w_1^k 自体の語としての結付きの強さを評価している．すなわち式 (10) による前接度が高い語は，語としてもまとまりが強く，さらに他の語を前接して結付く力も強いといえる．

3.4 後接度

ある語の直前に別の語を加えて新たな語を構成するという観点から，「後接度」(Suc) の尺度を以下で定義する．

$$Suc(w_1^k) = \sum_{w_{k+1} \in W} Con(w_1^k, w_{k+1}) = \sum_{w_{k+1} \in W} \mathcal{F}(w_1^k, w_{k+1}) \quad (12)$$

すなわち，語 w_1^k に後接して新たに単語を 1 つ加えて得られるすべての語について，結合度の総和をとった

値である．ここで前接度の場合と同様にして式 (12) を書き改めると以下の通りとなる．

$$\begin{aligned} &Suc(w_1^k) \\ &= P(w_1^k)\mathcal{D}\big(P(W_{k+1}|w_1^k)||P(W_{k+1})\big) \\ &\quad + \delta_s(w_1^k)\mathcal{F}(w_1^k) \end{aligned} \tag{13}$$

すなわち後接度が高い語は，語としてもまとまりが強く，さらに他の語を後接して結付く力も強いといえる．

3.5 文脈度

語が与えられた場合に，その語の前後に接続する語がどれくらい特定されるかを，「文脈度」(Cxt) として以下で定義する．

$$\begin{aligned} &Cxt(w_1^k) \\ &= Pre(w_1^k) + Suc(w_1^k) \\ &\quad - \big(\delta_p(w_1^k) + \delta_s(w_1^k)\big)\, Con(w_1^k) \\ &= P(w_1^k)\mathcal{D}(P(W_0|w_1^k)||P(W_0)) \\ &\quad + P(w_1^k)\mathcal{D}(P(W_{k+1}|w_1^k)||P(W_{k+1})) \end{aligned} \tag{14}$$

前後に接続する語への影響だけを評価するために，前接度と後接度の和から，その語自体を構成する単語どうしの結合度を差し引くものである．カルバックライプラー情報量の非負性から必ず $Cxt(w_1^k) \geq 0$ であり，等号が成立するのは w_1^k が前後に出現する語とはまったく独立に起きる場合である．これは具体的には「基礎・的・知見」「基本・構成・要素」等，(実験で用いた) コーパス中で他の単語とは複合せず常に単独で出現している語が相当している．

3.6 重要度

最後に，以上で定義した各尺度を総合的に評価する指標として「重要度」(Imp) を以下で定義する．

$$\begin{aligned} Imp(w_1^k) &= Cxt(w_1^k) + App(w_1^k) \\ &= Pre(w_1^k) + Suc(w_1^k) + App(w_1^k) \\ &\quad - \big(\delta_p(w_1^k) + \delta_s(w_1^k)\big)\, Con(w_1^k) \\ &= P(w_1^k)\mathcal{D}(P(W_0|w_1^k)||P(W_0)) \\ &\quad + P(w_1^k)\mathcal{D}(P(W_{k+1}|w_1^k)||P(W_{k+1})) \\ &\quad - P(w_1^k)logP(w_1^k) \end{aligned} \tag{15}$$

定義より必ず $Imp(w_1^k) > 0$ が成り立つ．ただし，上記の定義では，「語」が構成単語数にかかわらず単一の概念に対応するものとしている．これに対して，「語」は構成要素である単語それぞれが表す概念を複合したものであり，長い語ほど情報量が多いという立場からは，以下の定義も可能である．

$$\begin{aligned} Imp^*(w_1^k) &= Cxt(w_1^k) + Con(w_1^k) \\ &= Pre(w_1^k) + Suc(w_1^k) \\ &\quad - \big(\delta_p(w_1^k) + \delta_s(w_1^k) - 1\big)\, Con(w_1^k) \\ &= P(w_1^k)\mathcal{D}(P(W_0|w_1^k)||P(W_0)) \\ &\quad + P(w_1^k)\mathcal{D}(P(W_{k+1}|w_1^k)||P(W_{k+1})) \\ &\quad + \big(\delta_p(w_1^k) + \delta_s(w_1^k) - 1\big)\, \mathcal{F}(w_1^k) \end{aligned} \tag{16}$$

式 (15) と式 (16) の違いが顕著になるのは，つねにテキスト中で単独で出現する単語を評価する場合である．たとえば「明らか」という語について考えると，この語は通常他の語とは接続しないので，$Pre($明らか$) = Suc($明らか$) = 0$ となる．式 (15) による評価では $Imp($明らか$) = -P($明らか$)logP($明らか$) > 0$ となり，「明らか」が多用されるほど重要度は高く評価されるが，式 (16) による評価では $Imp^*($明らか$) = 0$ である．

Imp^* は高頻度で構成単語数が多く独立性が高い語に高い評価値を与えるという点で C-value に類似した尺度である．これまでの実験結果から経験的に，重要な単語は文脈度の値も大きくなる場合が多く，結合度の値はゼロであるものの Imp^* を用いても高い評価が得られることから，Imp^* の方が直観にあった結果が得られる場合が多いと考えている．

4 計算の実現法

4.1 セグメント集合の木構造表現

前節で定義した各尺度の依存関係を図 3 に示す．このように語 w_1^k に対する各尺度の値は，

(a) w_1^k の出現頻度

(b) w_1^k の FQ 値（結合度）

(c) その語の前後に単語を 1 つ追加して得られる各語の (a) と (b) の値

から求めることができるので，実装上は「語」を 1 つの「ノード」に対応させ，テキストから抽出したセグメント集合全体を 1 つの木構造で表現すると計算が容易である．

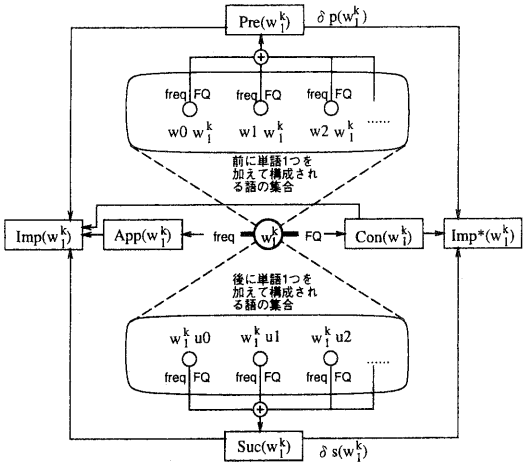

図 3: 各尺度の依存関係

具体的には，ノード n_i に対応する語を $Term(n_i)$，セグメント集合に含まれるすべての語の集合を $\mathcal{W}^*$ として，$Term(n_i) = w_1^k$ であるとき，n_i の下位ノードを以下で定義する．

$$Low(n_i) = \{n_j | Term(n_j) = (w_1^k, w_{k+1})\} \qquad (17)$$

ただし $w_{k+1} \in \mathcal{W}$，$Term(n_j) = (w_1^k, w_{k+1}) \in \mathcal{W}^*$ とする．また n_i の上位ノードを以下で定義する．

$$Upp(n_i) = \{n_j | Term(n_j) = (w_0, w_1^k)\} \qquad (18)$$

ただし $w_0 \in \mathcal{W}$，$Term(n_j) = (w_0, w_1^k) \in \mathcal{W}^*$ とする．ここで木構造における「下位ノード」と「上位ノード」は非対称の関係であることに注意が必要である．すなわち,「情報・検索」は「情報」の下位ノードであるが,「情報・検索」の上位ノードは「情報」ではなく,「画像・情報・検索」などとなる．

上記の定義のもと，$Term\ (*root*) = \emptyset$ なる特別な始点ノード $*root*$ から出発して順次下位ノードを展開し，最後に逆方向の木構造を作成する手順にしたがって，各ノードの上位ノードを求める．木構造を生成してしまえば，ノードごとの出現頻度および FQ 値に基づき，結合度，出現度，前接度，後接度，文脈度，重要度の値は直ちに計算できるので，各尺度に基づく語のランキングは容易である．

4.2 枝刈りのための上限下限値の計算

セグメント集合からの木構造生成は任意長の n グラム生成に対応しており，展開するノード数が多くなり過ぎる場合には,「枝刈り」を行う必要がある．このために以下で，FQ 値の上限および下限値を与える式を導く．

まず，式 (11) において，カルバックライプラー情報量の非負性から以下の不等式が導ける．

$$\sum_{w_0 \in \mathcal{W}} \mathcal{F}(w_0, w_1^k) \geq \delta_p(w_1^k)\mathcal{F}(w_1^k) \qquad (19)$$

等号が成立するのは，$P(w_0|w_1^k) = P(w_0)$，すなわち前接語が w_1^k とは独立に定まる場合である．現実的には，w_1^k が必ず語の先頭に現れる場合が相当している．式 (13) においても同様にして，カルバックライプラー情報量の非負性から以下の関係が導ける．

$$\sum_{w_{k+1} \in \mathcal{W}} \mathcal{F}(w_1^k, w_{k+1}) \geq \delta_s(w_1^k)\mathcal{F}(w_1^k) \qquad (20)$$

等号が成立するのは，$P(w_{k+1}|w_1^k) = P(w_{k+1})$，すなわち後接語が w_1^k とは独立に定まる（w_1^k が必ず語の末尾に現れる）場合である．

一方，上限値については次式が成立する．

$$\begin{aligned}
&\sum_{w_0 \in \mathcal{W}} \mathcal{F}(w_0, w_1^k) \\
&= \sum_{w_0 \in \mathcal{W}} P(w_0, w_1^k) log \frac{P(w_0, w_1^k)}{P(w_0)P(w_1)\cdots P(w_k)} \\
&\leq \sum_{w_0 \in \mathcal{W}} P(w_0, w_1^k) log \frac{P(w_1^k)}{P(w_0)P(w_1)\cdots P(w_k)} \\
&= \delta_p(w_1^k)P(w_1^k) log \frac{P(w_1^k)}{P(w_1)\cdots P(w_k)} \\
&\quad + \sum_{w_0 \in \mathcal{W}} P(w_0, w_1^k) log \frac{1}{P(w_0)} \\
&\leq \delta_p(w_1^k)\mathcal{F}(w_1^k) + \sum_{w_0 \in \mathcal{W}} P(w_0) log \frac{1}{P(w_0)} \\
&= \delta_p(w_1^k)\mathcal{F}(w_1^k) + H(W) \qquad (21)
\end{aligned}$$

ただし，$H(W) = -\sum_{w \in \mathcal{W}} P(w) log(P_w)$ は単語の生起に関する自己情報量とする．また不等号では $P(w_1^k)$，$P(w_0) \geq P(w_0, w_1^k)$ を用いた．同様に

$$\sum_{w_{k+1} \in \mathcal{W}} \mathcal{F}(w_1^k, w_{k+1}) \leq \delta_s(w_1^k)\mathcal{F}(w_1^k) + H(W) \qquad (22)$$

が成立する．式 (8)(9) および式 (19)–(22) より結局，

$$\begin{aligned}
\delta_p(w_1^k)Con(w_1^k) &\leq Pre(w_1^k) \\
&\leq \delta_p(w_1^k)Con(w_1^k) + H(W) \\
&\leq Con(w_1^k) + H(W) \qquad (23)
\end{aligned}$$

および

$$\begin{aligned}\delta_s(w_1^k)Con(w_1^k) &\leq Suc(w_1^k)\\ &\leq \delta_s(w_1^k)Con(w_1^k)+H(W)\\ &\leq Con(w_1^k)+H(W) \qquad (24)\end{aligned}$$

となり，$Con(w_1^k)$ により $Pre(w_1^k)$ および $Suc(w_1^k)$ の上限値と下限値が求まる．また個々の FQ 値について，

$$\begin{aligned}&\mathcal{F}(w_0,w_1^k)\\ &= P(w_0,w_1^k)log\frac{P(w_0,w_1^k)}{P(w_0)P(w_1)\cdots P(w_k)}\\ &\leq P(w_1^k)log\frac{P(w_1^k)}{P(w_1)\cdots P(w_k)}+P(w_0,w_1^k)log\frac{1}{P(w_0)}\\ &\leq \mathcal{F}(w_1^k)+min\left[P(w_0),P(w_1^k)\right]log\frac{1}{P(w_0)} \qquad (25)\\ &\leq \mathcal{F}(w_1^k)+P(w_1^k)logF \qquad (26)\end{aligned}$$

であり，同様に

$$\begin{aligned}&\mathcal{F}(w_1^k,w_{k+1})\\ &\leq \mathcal{F}(w_1^k)+min\left[P(w_1^k),P(w_{k+1})\right]log\frac{1}{P(w_{k+1})} \qquad (27)\\ &\leq \mathcal{F}(w_1^k)+P(w_1^k)logF \qquad (28)\end{aligned}$$

が成り立つ．ここで，式 (25)，式 (27) はそれぞれ，式 (26)，式 (28) よりも厳しい上限を与えるが，その計算には $P(w_0)$ または $P(w_{k+1})$ の値が必要である．また式 (26)，式 (28) で，最大語頻度 f_{max} が既知であるとして，総のべ語数 F を f_{max} で置換えてもよい．

以上の上限値および，$App(w_0,w_1^k), App(w_1^k,w_{k+1}) \leq App(w_1^k)$ を用いると，上位ノードおよび下位ノードに関する各尺度の値の上限値を求めることが可能で，これを利用して展開するノード数を削減できる．現在の実装では，その他に頻度や語数（すなわち n グラムそのもの）による足切りもサポートしている．しかし，完全な数え上げを保証したい場合には，ここで求めた上限値に基づく枝刈りが有効であると考えられる．

5 考察

現在，与えられたテキストからセグメント集合を切り出し，木構造表現に変換した上で計算を行うプログラムを試行実装し，各尺度に基づくランキングを図 4のインタフェースで閲覧可能にするとともに，用語抽出やテキスト分類タスクによる定量的な評価について検討を行っている．

詳細な評価は今後の課題であることから，ここでは予備的な実験として，国立情報学研究所の学会発表データベース [6] を対象として計算を行った結果を簡単に述べる．分析の対象としたテキストの総量は学術文献の題目および抄録で約 200M バイトである．形態素解析は Chasen Ver.2.02，セグメント抽出には以下のルールを用いた．現時点でこのルールは多分に経験的なものであり，たとえばテキストが学術論文であることから感動詞を名詞として扱うなどのヒュリスティックなルールを含んでいる．

MEISHI = {（非自立，代名詞，接尾を除く）名詞，未知語，感動詞 }
SHUSHOKU = {（連用テ，基本型を除く）形容詞，接頭詞-名詞接続，名詞-接尾 }
セグメント = { MEISHI | SHUSHOKU }* { MEISHI }

また分析結果を出力する際には，接尾辞で開始する語，接頭辞で終了する単語列，および，つねに特定の語の部分文字列として出現する語を除外した．このことから，テキスト原文中の誤りや英文字表記の場合を除き，出力語の大半は語として意味があるものになっている．

以上の条件のもとで，形態素解析の結果に基づき抽出したセグメント数は，のべで約 1.4×10^7 個，異なりで約 1.9×10^6 個，セグメントあたりの平均語数は 1.67 語，4 語以上の語を含むセグメントは約 9.2×10^4 個，全体の約 6%存在した．また，得られたセグメント集合から木構造を作成し，それぞれの尺度の数値を計算したところ，実行時間は Pentium 696 MHz の

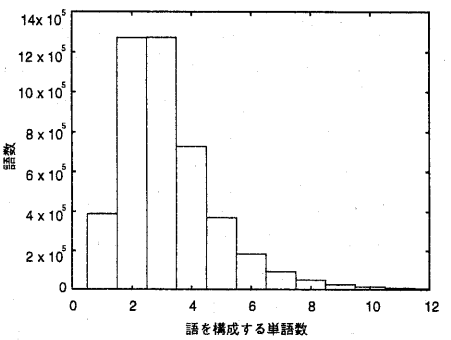

図 5: 語の長さ分布

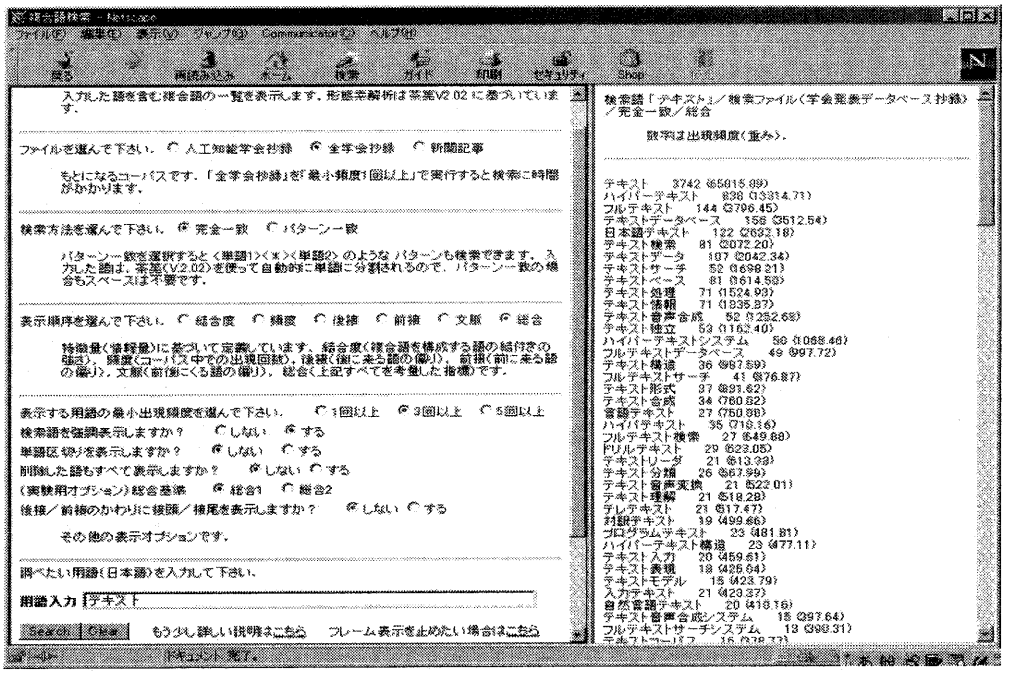

図 4: 検索結果画面

Linux 上で約 1 時間半であった．結果として出力された語の総数は約 4.4×10^6 語で，図 5に示した構成単語数の分布からわかるように，通常の n グラムでは対象としない 4 以上の単語からなる語も相当数含まれている．現在の実装は試行的なものであり，効率面では改善が必要とされるが，この実験で用いたテキストの規模であれば，特に枝刈りを行わなくても，使用に耐える処理速度ですべての語の数え上げを行うことが可能であった．

最後に，現在の実装で設定している上限値の設定はかなり緩いもので，枝刈りの効果は高々数割程度のノード数削減となっている．枝刈りの効果については，注目する尺度，コーパス規模や切り出したセグメントの長さ分布，メモリと処理時間のトレードオフ，ノード展開順序等，さまざまな要因が関係すると考えられることから，定量的な評価が今後の課題となっている．また，本稿では単純に $P(w_1^k) = freq(w_1^k)/F$ として議論を行ったが，実装上は，$P(w_1^k)$ の推定に確率的言語モデルによる各種スムージング法 [7] を用いることは容易である．ただし，この際の上限値計算については今後の検討課題となっている．

参考文献

[1] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸: 日本語形態素解析システム「茶筌」Version 2.0 使用説明書 第 2 版, NAIST Technical Report NAIST-IS-TR99012, 奈良先端科学技術大学院大学 (1999).

[2] Frantzi, K. T. and Ananiadou, S.: Extracting Nested Collocations, *Proc. of COLING'96*, pp. 41–46 (1996).

[3] Nakagawa, H. and Mori, T.: Nested Collocation and Compound Noun for Term Extraction, *Proc. of the First Workshop on Computational Terminology (COMPTERM'98)*, pp. 64–70 (1998).

[4] Aizawa, A.: The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures, *Proc. of ACM SIGIR2000*, pp. 104–111 (2000).

[5] 相澤彰子: 語と文書の共起に基づく特徴度の数量的表現について, 情報処理学会論文誌, Vol. 41, No. 12, pp. 3332–3343 (2000).

[6] NACSIS(ed.): *NTCIR Workshop 1 - proc. of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, National Center for Science Information Systems (NACSIS), Japan (1999).

[7] 北研二: 確率的言語モデル, 東京大学出版会, 東京 (1999).