特徴語抽出のための数量尺度の定義と適用例

相澤 彰子

国立情報学研究所

akiko@nii.ac.jp

1　はじめに

形態素解析の結果を利用したテキストからのキーワード抽出では，(1) 形態素解析による単語をそのまま「語」として用いる場合（最短一致），(2) 形態素情報を手がかりに，語を構成する最長の単語列を取出す場合（最長一致）のいずれかが用いられることが多い．しかし，語が k（>3）個の単語から構成される場合には，その他にも多数の部分単語列を考えることが可能である．たとえば「自然言語処理」は「自然」「言語」「処理」「自然言語処理」以外に「自然言語」「言語処理」という部分単語列を含んでいる．これら中間的な単語のまとまりが，用語あるいはテキストのキーワードとして役に立つものであるか否かについては，実際に用語抽出を行った上で，定量的な評価を行うことが有用であると考えられる．

しかし現実に，このような定量的な評価を行おうとすると，以下の2つの困難が生じる．まず，(1) 構成単語数の異なる単語列を数値的に比較することが必要になるが，2単語の共起に基づく従来の統計尺度はこのような比較を想定していない．また，(2) 部分単語列は kの数に対して組合せ的な数だけ存在することから，大量のテキストを対象とする場合はすべての単語列を数え上げることがむずかしい．そこで本稿では，(1)(2) に対応可能な数量尺度の定義と計算方法に焦点をあてて，現在の検討結果を報告する．

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．本稿で想定する語の抽出手順は，上記のようにごく単純なものであり，用語抽出を行う際の前段階としてまず，(1) 構成語数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点を主たる目的としている．

以下，**2.** および **3.** でこれら2つの問題に関する現在のアプローチを述べる．また **4.** で簡単な適用例を示し，**5.** で今後の課題を検討する．

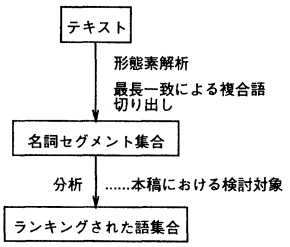

図 1: 想定する処理の流れ

2　共起関係抽出に関する統計尺度

2.1　FQ 値の定義式

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．

本稿では，任意の k 語の共起に関する基本的な尺度として，確率と情報量の席である「FQ 値」(Feature Quantity Value) [1][2] を用いる．具体的には語 w_1^k に対する FQ 値を次式で定義する．

$$\mathcal{F}(w_1^k) = P(w_1^k) log \frac{P(w_1^k)}{P(w_1) \cdots P(w_k)} \qquad (1)$$

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．

$$P(w_1^k) = \frac{freq(w_1^k)}{F} \quad (2)$$

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，

$$P(w_1^k) = P(\emptyset, w_1^k) + \sum_{w_0 \in \mathcal{W}} P(w_0, w_1^k) \quad (3)$$

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，

$$P(w_1^k) = P(w_1^k, \emptyset) + \sum_{w_{k+1} \in \mathcal{W}} P(w_1^k, w_{k+1}) \quad (4)$$

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，$\delta_p(w_1^k) = \frac{P(w_1^k) - P(\emptyset, w_1^k)}{P(w_1^k)}$ および $\delta_s(w_1^k) = \frac{P(w_1^k) - P(w_1^k, \emptyset)}{P(w_1^k)}$ とする．定義より明らかに，$0 \leq \delta_p(w_1^k) \leq 1$，$0 \leq \delta_s(w_1^k) \leq 1$ である．

2.2 FQ値に基づく異なる尺度の定義

(a) 結合度

式(1)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していることから，語を構成する単語の共起の度合いを示す「結合度」(Con) を以下で定義する．

$$Con(w_1^k) = \mathcal{F}(w_1^k) = P(w_1^k) log \frac{P(w_1^k)}{P(w_1) \cdots P(w_k)} \quad (5)$$

式(5)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．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），結合度は大きくなる．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，結合度はゼロである．

(b) 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．

$$App(w_1^k) = -P(w_1^k) log P(w_1^k) \quad (6)$$

ここで $-logP(w_1^k)$ は w_1^k の生起による情報量である．式(6)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(6)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．

(c) 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) を以下で定義する．

$$Pre(w_1^k) = \sum_{w_0 \in W} Con(w_0, w_1^k) \quad (7)$$

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(7)は以下の形に書き改めることができる．

$$Pre(w_1^k) = \mathcal{D}\big(P(W_0|w_1^k) || P(W_0)\big) + \delta_p(w_1^k)\mathcal{F}(w_1^k) \quad (8)$$

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．

式(8)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(7)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．

(d) 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) を以下で定義する．

$$Suc(w_1^k) = \sum_{w_{k+1} \in W} Con(w_1^k, w_{k+1}) \quad (9)$$

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 (9) を書き改めると以下の通りとなる．

$$Suc(w_1^k) = P(w_1^k)\mathcal{D}\big(P(W_{k+1}|w_1^k)||P(W_{k+1})\big)+\delta_s(w_1^k)\mathcal{F}(w_1^k) \tag{10}$$

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．

(e) 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．

$$\begin{aligned} Cxt(w_1^k) &= Pre(w_1^k) + Suc(w_1^k) \\ &\quad - \big(\delta_p(w_1^k) + \delta_s(w_1^k)\big)\, Con(w_1^k) \\ &= P(w_1^k)\mathcal{D}(P(W_0|w_1^k)||P(W_0)) \\ &\quad + P(w_1^k)\mathcal{D}(P(W_{k+1}|w_1^k)||P(W_{k+1})) \end{aligned} \tag{11}$$

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．

(f) 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．

$$Imp(w_1^k) = Cxt(w_1^k) + App(w_1^k) \tag{12}$$

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．

$$Imp^*(w_1^k) = Cxt(w_1^k) + Con(w_1^k) \tag{13}$$

式 (12) と式 (13) 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 (12) 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 (13) 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[3] に類似した尺度である．

3 計算の実現法

3.1 セグメント集合の木構造表現

語 w_1^k が与えられる場合に，前節で定義した各尺度の値は，

(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（双方向の）木構造で表現すると計算が容易である．

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．

$$Low(n_i) = \{n_j | Term(n_j) = (w_1^k, w_{k+1})\} \tag{14}$$

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．

$$Upp(n_i) = \{n_j | Term(n_j) = (w_0, w_1^k)\} \tag{15}$$

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．

3.2 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．簡単な計算により，以下の不等式が導ける．

$$Pre(w_1^k), Suc(w_1^k) \leq Con(w_1^k) + H(W) \quad (16)$$

ただし，$H(W) = -\sum_w P(w)log(w)$ は単語の自己情報量の期待値である．また個々の値について，

$$Con(w_0, w_1^k), Con(w_1^k, w_{k+1}) \leq Con(w_1^k) + P(w_1^k)logF \quad (17)$$

が成り立つ．さらに，定義より明らかに

$$App(w_0, w_1^k), App(w_1^k, w_{k+1}) \leq App(w_1^k) \quad (18)$$

である．以上の上限値およびカルバックライブラー情報量の非負性を用いると，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，伝統的な頻度や語数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．

4 適用例

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Web上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．これらの分析結果については別途報告することとし，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 [4] 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，結果として出力された語の総数は約 4.4×10^6 語であった．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．

5 今後の課題

用語抽出タスクである TMREC[4] の正解集合を用いた実験によると，本稿で定義した尺度のうちで結合度がもっともよい結果を示しており，2 個以上の語から構成される複合語に注目すると，用語抽出の尺度として人間の直観にも比較的合致したランキングが得られる．ただし特に語が 1 単語から構成される場合には，いずれの尺度を用いても，分野特有の表現と一般語の区別に必ずしも成功するわけではない．本稿で定義した尺度を単一分野のコーパスに適用して抽出されるのは，あくまで語としてまとまりのよい単語列であり，これらの単語のまとまりが当該分野を特徴づける用語であるか否かについては，異なる観点からの評価が必要である．具体的には，他分野のコーパスにおける抽出結果と対比することで，よいランキングが得られることを予備実験において観察している．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，$P(w_1^k)$ の推定には確率的言語モデルによるスムージング手法を適用することの有効性も観察しており，この際の上限値の計算が今後の検討課題となっている．

参考文献

[1] Akiko Aizawa: "The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures," Proc. of ACM SIGIR2000, pp.104–111 (2000).
[2] 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, 41(12), pp.3332–3343 (2000).
[3]K. T. Frantzi and S. Ananiadou: "Extracting Nested Collocations," Proc. of COLING'96, pp.41-46 (1996).
[4] NACSIS: "NTCIR Workshop 1 - proc. of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition," (1999).