社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2003－FI－ 72　(10)
2003－NL－157　(10)
2003／9／30

複数尺度の統計的統合法とその専門用語抽出への応用

内山将夫 井佐原均
通信総合研究所
{mutiyama,isahara}@crl.go.jp

本稿では，複数尺度が与えられたとき，それらを教師なしで統合して１つの尺度を構成する統計的な方法を提案する．その方法は，尺度値自体ではなく，尺度値によりソートされた標本における累積経験確率を利用して，複数尺度を統合する．提案手法の効果を調べるために，専門用語の抽出実験をした．その結果，提案手法は，単に複数尺度の値を乗算するよりも性能が高かった．また，その統合の効果は，代表的な教師あり学習である Support Vector Machine を利用した場合と同程度であった．

Combination of Multiple Measures and its Application to Term Extraction

Masao Utiyama and Hitoshi Isahara
Communications Research Laboratory
{mutiyama,isahara}@crl.go.jp

This paper discusses methods of combining multiple measures through statistical analysis. These methods use cumulative probabilities obtained by sorting samples with given measures, instead of using their values, to combine them. The proposed methods were applied to technical term extraction. The experiments demonstrated that the proposed methods were more effective than simple multiplication. It also revealed that the methods were as effective as support vector machines, a representative machine learning method.

1　はじめに

ある集合$X = \{x_1, x_2, \ldots, x_m\}$について，$X$の要素には,「良さ」により，全順序関係が定義できるとする．そして，ある関数gがあり，以下が成立するとする．

$$g(x_i) < g(x_j) \Longleftrightarrow x_i \text{よりも} x_j \text{が良い}$$

$$g(x_i) = g(x_j) \Longleftrightarrow x_i \text{と} x_j \text{が同等に良い}$$

我々の目的は，このgを推定することである．

ここで，我々の手元には，$f_1, f_2, \ldots, f_n$というn個の関数(尺度)があり，それぞれ，gを近似しているとする．つまり，f_iについて，$f_i(x_j) < f_i(x_k)$ ならば $g(x_j) < g(x_k)$ が，ある程度は成立しているとする．ここで，ある程度成立しているとは，たとえば，Xをgにより順位付けた場合とf_iにより順序付けた場合とで，それらの間の順位相関がある程度は高いということである．

このようなn個のf_iから，gを推定するのが，我々の目的であり，本稿で対象とする問題設定である．

gを推定するときには，教師あり学習と教師なし学習の2通りがある．まず，教師あり学習だが，これは，ある訓練データXについて，$x \in X$について，$g(x)$と$f_i(x)$とが与えられていて，それらを利用して，f_iからgを推定する方法である．この場合には，たとえば，$f(x) = \sum_{i=1}^{n} a_i f_i(x)$なる$f(x)$を定義し，これと$g(x)$とがなるべく近くなるように$a_i$を推定することにより，$g$を推定する．その他にも，教師あり学習には，さまざまな，機械学習の手法がある．

次に，教師なし学習だが，これは，gが未知である．我々の手元には，f_iしかない．そのため，(Xと)f_iのみにより，gを推定しないといけない．そのための方法としては，単純に，和をとり，$f(x) = \sum_{i=1}^{n} f_i(x)$としたり，あるいは，積をとり，$f(x) = \prod_{i=1}^{n} f_i(x)$としたりする場合がある．このとき，$f_i(x)$に適当な変換を掛ける場合もある．

この二つの学習のうちで，我々が対象とするのは，教師なし学習の方である．我々は，(Xと)f_iのみが与えられている状態から，gを推定する方法を，本稿で提案する．そして，専門用語抽出という適用対象について，提案手法と，単純な和や積による尺度統合法とを比較し，我々の方法の有効性を示す．

2　複数尺度の統合法

我々の方法は，経験確率に基づいたものであり，簡明なものである．すなわち，以下の $F_{cum}(x)$あるいは$F_{mul}(x)$をもって，$g(x)$の推定値とする．ここで，$F_{cum}(x)$と$F_{mul}(x)$とは，尺度数nが1のときには等価であるが，$n \geq 2$のときには，統合方法に違いがでる．なお, 4節の実験では，$F_{mul}(x)$の方が精度が若干良く，かつ，計算量も少ないので，$F_{mul}(x)$を使う方が良いようであるが，これらは関連しているので，2つとも説明する．

まず，用語を定義する．我々は，専門用語の抽出を対象にしているので，それを念頭に置いて用語を定義する．

$X =$ 尺度値を割当てたいような単語や複合語の集合

$$N(x) = x(\in X) \text{ の出現頻度}$$

$$N = \sum_{x \in X} N(x)$$

$$R(x) = N(x)/N$$

$$[f_i(x') \leq f_i(x)] = \begin{cases} 1 & \text{if } f_i(x') \leq f_i(x) \\ 0 & \text{otherwise} \end{cases}$$

$R(x)$はxの経験確率である．このとき，

$$F_{cum}(x) = \sum_{x' \in X} R(x') \prod_{i=1}^{n} [f_i(x') \leq f_i(x)] \qquad (1)$$

である．ここで，$\prod_{i=1}^{n}[f_i(x') \leq f_i(x)]$は，全ての$f_i$について，$f_i(x') \leq f_i(x)$のときに，1 となり，それ以外では 0 となる．(1) 式は，その値が1となるようなx'についてのみ，$R(x')$を足した値である．

また，尺度iのみについて，(1) 式と同様に

$$F^i_{cum}(x) = \sum_{x' \in X} R(x')[f_i(x') \leq f_i(x)] \qquad (2)$$

とすると，F_{mul}は以下である．

$$F_{mul}(x) = \prod_{i=1}^{n} F^i_{cum}(x) \qquad (3)$$

$F^i_{cum}(x)$は，尺度iのみについての，$f_i(x)$以下の尺度値の要素の確率の和である．$F_{mul}(x)$は，それらを全尺度について掛けたものである．

以下では，F_{cum}とF_{mul}とを例により説明するが，その前に，2要素xとx'の比較のための用語を定義する．まず，$x \equiv x'$とは，全てのf_iについて，$f_i(x) = f_i(x')$であり，$x \prec x'$とは，全てのf_iについて，$f_i(x) < f_i(x')$であり，$x \preceq x'$とは，全てのf_iについて，$f_i(x) \leq f_i(x')$であると定義する．そして，$x \preceq x'$か$x' \preceq x$ のいずれかが成立するようなxとx'とは「比較可能」であると言い，比較可能でないときには,「比較不能」であると言う．また，$x \preceq x'$であるとき，xはx'より「劣位」であり，x'はxより「優位」であると言う．

$F_{cum}(x)$とは，xより劣位なものについての経験確率の和である．また，$F_{mul}(x)$は，周辺分布の確率の積により，$F_{cum}(x)$と同様なものを計算しているものである．

$F_{cum}(x)$と$F_{mul}(x)$は，1次元における経験確率分布を利用した統計的検定を多次元に拡張する一つの方法である．

これらの統合法の性質をみるために，まず，$n=1$，つまり，尺度が1つの場合をみてみる．このとき，

$$F(x) = F_{cum}(x) = F_{mul}(x) = \sum_{x' \in X} R(x)[f_1(x') \leq f_1(x)]$$

である．この場合，Xの要素を適当に並べかえて，$i \leq j$なら，$F(x_i) \leq F(x_j)$となるようにできて，そのときの$f_1(x)$と$F(x)$との関係は図1のようになる．

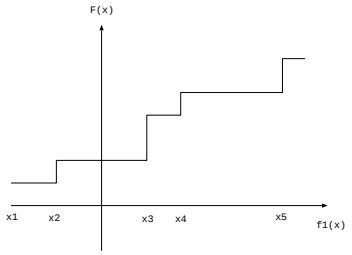

図 1: $f_1(x)$と$F(x)$との関係.

図1より分かるように，$F(x)$は，$f_1(x)$の階段関数である．このとき，$F(x)$は，xより劣位なるものの累積経験確率となっている．したがって，たとえば，$F(x)=0.95$のときには，$x \prec x'$なるx'の出現確率は$\sum_{x \prec x'} \Pr(x') = 1-F(x) = 0.05$である．つまり，$x$より，尺度$f_1$の観点から「良い」要素$x'$が出現する確率は0.05である．すなわち，尺度$f_1$の観点から，$x$より良いものが出現する確率は有意水準5%である．

以上から分かるように，$n=1$の場合には，$F(x)$は，Xの確率分布を経験確率分布とした場合における，統計的検定による有意水準と等価である．$F(x)$が大きい値となるのは，$x' \prec x$なるx'が多いときである．

また，$n=1$のときには，

$$F(x) \leq F(x') \Longleftrightarrow f_1(x) \leq f_1(x')$$

が成立している．したがって，$n=1$のときには，$F(x)$と$f_1(x)$とは順序尺度の観点からは同等なので，$F(x)$，すなわち，$F_{cum}(x)$あるいは$F_{mul}(x)$を$f_1(x)$の代りに利用することは問題ない．

次に，$n=2$の場合を考える．まず，図2のように，x_1とx_2とが比較可能な場合を考える．このとき，図中の○によりXの要素を指すとすると，$x_1 \preceq x_2$であるので，$F_{cum}(x_1) \leq F_{cum}(x_2)$，かつ，$F_{mul}(x_1) \leq F_{mul}(x_2)$である．これは，$i=1,2$について，$f_i(x_1) \leq f_i(x_2)$であるので，こうなって当然であり，問題はない．

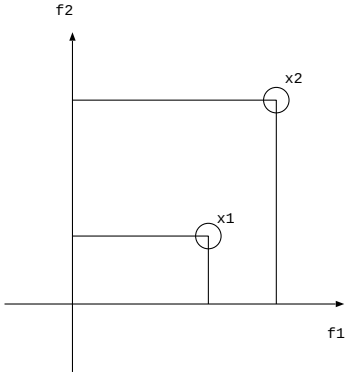

図 2: x_1とx_2とが比較可能な場合.

次に，図3のように，x_1とx_2とが比較不能な場合を考える．このとき，$f_1(x_1) < f_1(x_2)$かつ$f_2(x_1) > f_2(x_2)$である．このように，尺度間において，二つの対象に与える尺度値の順位に矛盾がある場合にも，F_{cum}やF_{mul}を利用すれば，一貫した順位を付けることができる．たとえば，図3には，x_1,x_2の他に，7点が出現し，それぞれの頻度が全て1回ずつだとすると，$F_{cum}(x_1) = \frac{5}{9}$，$F_{cum}(x_2) = \frac{3}{9}$である．したがって，$F_{cum}(x_1) > F_{cum}(x_2)$であるので，$F_{cum}$の観点からは，$x_1$の方が「良い」要素である．同様に，$F_{mul}$については，$F^1_{cum}(x_1) = \frac{6}{9}$，$F^1_{cum}(x_2) = \frac{8}{9}$であり，$F^2_{cum}(x_1) = \frac{8}{9}$,$F^2_{cum}(x_2) = \frac{4}{9}$であるので$F_{mul}(x_1) = \frac{6}{9} \times \frac{8}{9} = \frac{48}{81}$, $F_{mul}(x_2) = \frac{8}{9} \times \frac{4}{9} = \frac{32}{81}$である．よって，$F_{mul}$の観点からも，$x_1$の方が「良い」要素である．

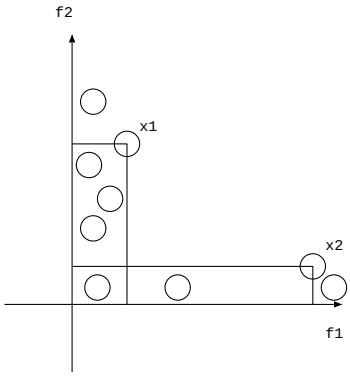

図 3: x_1とx_2とが比較不能な場合.

ここで，$n=2$の場合について，F_{cum}とF_{mul}の関係を示すと図4のようになる．図4では，$F_{cum}(x)$は実際に観測した場合の，xより劣位なものについての経験確

率の和である．また，$F^1_{cum}(x)$ と $F^2_{cum}(x)$ とは，周辺分布 $f_1(x)$ と $f_2(x)$ とについての確率である．したがって，$F_{mul}(x) = F^1_{cum}(x) \times F^2_{cum}(x)$ は，$F_{cum}(x)$ の (実測値でなく) 期待値である．

	$f_2(x)$ 以下	$f_2(x)$ より大	
$f_1(x)$ 以下	$F_{cum}(x)$		$F^1_{cum}(x)$
$f_1(x)$ より大			
	$F^2_{cum}(x)$		1

図 4: F_{cum} と F_{mul} との関係.

このように F_{cum} と F_{mul} とには密接な関係がある．そして，専門用語抽出に，これらを利用した場合の性能については，4節で述べる．ここでは，計算量について述べると，F_{mul} の計算は，各尺度ごとに，要素をソートすれば，それを利用して，容易に，F^i_{cum} を計算できるので，あとは，それを掛けるだけで良いので，m を要素数とすると，尺度の数は定数なので無視するとして，$O(m\log(m))$ である．一方，F_{cum} については，本稿では，単純に，各要素について，それ以下のものを数えあげるという方法をとったので，$O(m^2)$ である．なお，これについては，もっと改善する方法があると思われる．

これまでに，$n = 1, 2$ の場合をみたが，$n \geq 3$ についても，同様に F_{cum} や F_{mul} を計算できる．これらは，$n = 1$ の場合には，経験確率分布を利用した統計的検定に相当することをみた．$n \geq 2$ については，我々は，F_{cum} や F_{mul} は，経験確率分布を利用した統計的検定を，多次元に拡張したとみなせると考えている．

最後に，F_{cum} や F_{mul} が，どのように f_i を利用しているかについて述べると，これらは，f_i を順序尺度として利用している．したがって，任意の 2 要素間の大小関係が分かるような尺度であれば，f_i として利用できる[1]．

3 比較する統合方法

比較の対象とした統合方法は，F_{cum} と F_{mul} の他には，F_{sum}，Raw_{sum}，Raw_{mul}，$Norm_{sum}$ で，その定義は，それぞれ，$F_{sum}(x) = \sum_{i=1}^{n} F^i_{cum}(x)$, $Raw_{sum}(x) = \sum_{i=1}^{n} f_i(x)$, $Raw_{mul}(x) = \prod_{i=1}^{n} f_i(x)$, $Norm_{sum}(x) = \sum_{i=1}^{n} \frac{f_i(x)}{SD(f_i(X))}$．ただし，$SD(f_i(X))$ は，X の要素に対して f_i を適用した場合の標準偏差である．なお，$Norm_{mul}$ も同様に定義できるが，これと Raw_{mul} とは順序尺度としては等価である．

これらの統合法については，予備実験の結果から，F_{cum} と F_{mul} とは，それら以外の統合法全てよりも高精度であった．そのため，以下では，従来研究との比較という観点から，Raw_{mul} のみを追加し，全部で，F_{cum} と F_{mul} と Raw_{mul} の 3 つを比較する．

4 実験

本節では，各統合法を専門用語抽出に適用した結果について述べる．まず，(Frantzi and Ananiadou 1996, 1999; 中川, 湯本, 森 2003) を参照し，それらに使われている尺度から，それらの構成尺度を取り出す．次に，構成尺度の統合結果を統合法ごとに比較する．

本稿では，これら尺度の言語上の根拠等については，考察の対象外である．ただ，統合のための構成尺度としてのみ，これらを利用する．これらの尺度の根拠等については，先行研究を参照のこと．なお，尺度の表記法は，先行研究と若干異なる部分もある．

4.1 先行研究における尺度および本稿での構成尺度

(Frantzi and Ananiadou 1996) では，次の尺度を提案している．

$$\text{C-value}(CW) = (\text{Len}(CW) - 1) \times NTC(CW) \quad (4)$$

$$NTC(CW) = n(CW) - \frac{t(CW)}{c(CW)}$$

ただし，$0/0 = 0$ とする．また，

CW	専門用語の候補となる複合名詞
$\text{Len}(CW)$	CW の長さ (文字数)
$n(CW)$	コーパスにおける CW の出現頻度
$t(CW)$	CW を含むより長い複合名詞の出現頻度
$c(CW)$	CW を含むより長い複合名詞の異なり数

である[2]．ところが，(4) 式では，複合名詞の長さが 1 の場合には，尺度値が 0 となってしまうので，(中川他 2003) にならって，以下のように変更する．

$$\text{MC-value}(CW) = \text{Len}(CW) \times NTC(CW) \quad (5)$$

[1] F_{mul} および次節で定義する F_{sum} は，順序しか利用していない．その結果として，もし，全ての要素が 1 回しか出現しなく，かつ，要素間の尺度値に同一のものがない場合には，要素の順位のみを利用した場合と同等になる．すなわち，F_{mul} の場合には，各尺度で降順にソートした結果の順位を乗算して，その結果の値が小さいものほど良い要素とする，また，F_{sum} では同様なことを加算で行なう，ということと同等である．したがって，F_{mul} や F_{sum} は，順位を利用した尺度の統合を一般化したものと考えることができる．

[2] Len としては，文字数でなく，構成単語の数を取ることも考えられる．また，(Frantzi and Ananiadou 1999) と同様，Len ではなく，その対数を利用することも考えられる．しかし，予備実験では，長さを文字数で取り，かつ，Len のままのものが一番精度が高かった．

次に，(中川他 2003)で提案されている尺度について述べる．(中川他 2003) では次の尺度を提案している．

$$FLR_{\text{連接頻度}}(CW) = \text{Freq}(CW) \times LR_{\text{連接頻度}}(CW) \quad (6)$$

ここで，Freq(CW)は，候補となる複合名詞が単独で，他の複合名詞に包含されることなく出現した回数である．また，$LR_{\text{連接頻度}}(CW)$は，CWがL個の単語$W_1, W_2, \ldots W_L$からなるとしたとき，

$$LR_{\text{連接頻度}}(CW) = \left(\prod_{i=1}^{L}(FL_{\text{連接頻度}}(W_i)+1)(FR_{\text{連接頻度}}(W_i)+1)\right)^{\frac{1}{2L}}$$

ただし，$FL_{\text{連接頻度}}(W_i)$と$FR_{\text{連接頻度}}(W_i)$は，コーパス中の複合名詞全てをとりだしたとき，それぞれ，単語W_iの左側と右側に出現する全単語の頻度の和である．これを左右の連接頻度と呼ぶ．

よって，$LR_{\text{連接頻度}}(CW)$は，各構成単語W_iについて，左右の連接頻度に1を加えたものを乗算したものについて，幾何平均をとったものである．同様に，コーパス中の複合名詞全てをとりだしたとき，W_iの左右に出現する全単語の異り数を，それぞれ数えて，これを左右の連接種類数と呼び，$FL_{\text{連接種類数}}(W_i)$と$FR_{\text{連接種類数}}(W_i)$と記す．$LR_{\text{連接種類数}}(CW)$は以下のように定義される．

$$LR_{\text{連接種類数}}(CW) = \left(\prod_{i=1}^{L}(FL_{\text{連接種類数}}(W_i)+1)(FR_{\text{連接種類数}}(W_i)+1)\right)^{\frac{1}{2L}}$$

これを利用して，

$$FLR_{\text{連接種類数}}(CW) = \text{Freq}(CW) \times LR_{\text{連接種類数}}(CW) \quad (7)$$

も定義する．

(中川他 2003)では，彼等が比較した，MC-valueを含む尺度のうちでは，$FLR_{\text{連接頻度}}(CW)$がもっとも優れていると述べている．なお，彼等は，$FLR_{\text{連接種類数}}(CW)$については，精度評価をしていない．

以上，先行研究を参考にして，MC-value, $FLR_{\text{連接頻度}}$, $FLR_{\text{連接種類数}}$を定義した．これらの構成尺度は，次の5つである．Len, NTC, Freq, $LR_{\text{連接頻度}}$, $LR_{\text{連接種類数}}$．なお，以下では，$LR_{\text{連接頻度}}$をLR_nと記し，$LR_{\text{連接種類数}}$をLR_dと記す．これは，FLRやFLやFRについても同様である．

ここで，5つの構成尺度のうちで，LR_nとLR_dとは，FL_nやFL_dなどをRaw_{mul}により統合したものの幾何平均をとったものと考えることができる．そこで，それに対応するものとして，F_{cum}やF_{mul}で統合したものを考え，それらを$cumLR_n$, $cumLR_d$, $mulLR_n$, $mulLR_d$とする．具体的には，たとえば，$cumLR_n$については，まず，W_iのスコアを$NLR(W_i)$として，(1)式で，$f_1(W_i) = FL_n(W_i)$, $f_2(W_i) = FR_n(W_i)$として適用する．つまり，W'を単語として，

$$NLR(W_i) = \sum_{W'} R(W') \prod_{j=1}^{2}[f_j(W') \leq f_j(W_i)]$$

としたとき，その幾何平均をもって$cumLR_n$とする．

$$cumLR_n(CW) = \left(\prod_{i=1}^{L} NLR(W_i)\right)^{\frac{1}{L}} \quad (8)$$

その他についても同様である．

結局，構成尺度は次のものである．まず，全体に共通なものとして，Freq, Len, NTCの3つがあり，次に，Raw_{mul}のために，LR_n, LR_dがあり，F_{cum}用に，$cumLR_n$, $cumLR_d$があり，F_{mul}用に，$mulLR_n$, $mulLR_d$である．そのため，Raw_{mul}, F_{cum}, F_{mul}の各統合法ごとに5つの構成尺度があることになる．

比較の手順としては，まずは，単独の構成尺度の精度評価をし，次に，任意の2つの尺度について，それらを統合した場合を比較する．また，3,4,5個の尺度を使った場合について，網羅的に精度を評価する．更に，教師あり学習との比較もする．

4.2 実験材料

我々は，NTCIR-1のTMRECタスクで利用されたテストコレクション (Kageura 1999)を利用して，各種統合法を比較した．TMRECでは，NACSIS学術会議データベースから収集された人工知能の分野の1870の抄録をコーパスとして，そこから，専門用語を抽出するというタスクをした．このうち，主催者側で用意した参照用の用語リストは8834語である．この用語リストを正解用語リストと呼ぶことにし，以下では，この用語リストとの一致の度合が高い程精度が高いとみなす．このコーパスから用語候補を取り出すために，我々は，東京大学中川研究室・横浜国立大学森研究室で開発された用語抽出システム[3] を使用した．すなわち，茶筌[4]による形態素解析結果をその用語システムに入力し，その結果として，システムが用語候補として出力した17087語を利用

[3] http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/resource/termext/atr.html
[4] http://chasen.aist-nara.ac.jp/

した[5]．この用語候補のなかで正解用語リストに含まれるものは7024個である．

我々は，この用語候補リストを各種尺度でソートし，そのソートの結果，上位に正解用語が多いものほど良い尺度であると判断することにした．その評価の指標としては，平均精度(Average Precision, AP)を利用した．平均精度は，情報検索の評価(Baeza-Yates and Ribeiro-Neto 1999)や連語抽出の評価(Schone and Jurafsky 2001)にも使われていて，上位に正解が多いほど高い値を取るので，今回の評価の指標として適切である．ここで

$$\text{平均精度} = \frac{1}{K}\sum_{i=1}^{K}\frac{i}{H_i}$$

ただし，Kは，用語候補リスト中の正解用語数(7024)であり，$\frac{i}{H_i}$は，i番目の正解が抽出されたときの適合率であり，H_iは，i番目の正解が抽出されたときの，ソートされた用語候補における順位である．なお，以下で，単に精度と言うときには，平均精度のことであるとする．

4.3 実験結果

構成尺度の精度評価

まず，単独の構成尺度の平均精度を表1に示す．これより，頻度(Freq)よりも優れた尺度は，F_{cum}もしくはF_{mul}により統合された尺度のみである．一方，Raw_{mul}で統合されたLR_nやLR_dは頻度よりも平均精度が低い．これは，(中川他 2003)における観察と一致する．これより，F_{cum}やF_{mul}を利用することにより，単独の尺度でも，Freqよりも平均精度が高い尺度が構成できることが分った．

表 1: 構成尺度の平均精度.

尺度	平均精度
$cumLR_n$	0.541
$mulLR_n$	0.539
$cumLR_d$	0.536
Freq	0.535
$mulLR_d$	0.532
LR_n	0.532
LR_d	0.522
NTC	0.515
Len	0.425

2つの尺度の統合の精度評価

次に，F_{mul}，F_{cum}，Raw_{mul}のそれぞれで，各統合法についての，任意の2つの尺度を統合した結果について，それぞれの平均精度の降順にソートしたものの上位15位を表2に示す．また，参考のために，先行研究での尺度との比較を表3に示す．なお，これらの表および以下で，F_{mul}などの列にあるLR_nなどは$mulLR_n$などのことである．

表 2: 上位15位の平均精度.

F_{mul}	AP	F_{cum}	AP	Raw_{mul}	AP
Freq, Len	0.621				
		Freq, LR_n	0.616		
NTC, Len	0.615				
Freq, LR_n	0.613				
		Freq, LR_d	0.612		
Freq, LR_d	0.608				
NTC, LR_n	0.598				
		NTC, LR_n	0.598		
		Freq, Len	0.597		
		NTC, Len	0.595		
				Freq, LR_n	0.594
		NTC, LR_d	0.594		
NTC, LR_d	0.594				
				Freq, LR_d	0.593
				NTC, LR_n	0.580

表 3: 先行研究における尺度との比較.

統合法	構成尺度	AP	従来尺度
F_{mul}	Freq, Len	0.621	
F_{cum}	Freq, LR_n	0.616	
Raw_{mul}	Freq, LR_n	0.594	$FLR_{\text{連接頻度}}$
Raw_{mul}	NTC, Len	0.545	MC-value
Freqのみ		0.535	

表3より，各統合法の最高精度同士を比べると，F_{mul}とF_{cum}との精度差は，それほどない(1ポイント未満)が，Raw_{mul}とF_{mul}とは，2.7ポイントあるので，実質的な差であると言える．これより，統合法による精度の向上は顕著なものであると言える．この差は，3つ以上の構成尺度を考えたときには，もっと大きくなる．

3,4,5個の組み合せの精度評価

ここでは，3,4,5個の組み合わせについて，網羅的に統合法を適用し，それらを平均精度の降順にソートした結果を表4に示す．表4には，2尺度の統合結果における最高精度のものも載せてある．

表4より，重要な構成尺度を考えると，まず，FreqとLenとは，全統合法において，上位4位までが，必ず，この2つを含むことから，最重要であると言える．次に，上位を見てみると，LR_nかLR_dのどちらかを含むので，これらも重要である．しかし，これら2つを共に含むものは，かえって精度が低下する場合もある．この理由は，LR_nとLR_dとが独立でない[6]ためではないかと考えられ

[5]この用語抽出システムが出力する候補は形態素が連結されたものであるが，我々は，形態素が分割されたままの状態の用語候補を利用した．なお，本稿で比較したLR_nやLR_dは，我々の実装である．

[6]順位相関(Kendall's τ)は，$mulLR_n$と$mulLR_d$，$cumLR_n$と

る．なお，LR_n と LR_d では，LR_n の方が重要な尺度であるようである．これは，(中川他 2003) における観察と一致する．また，NTC も精度向上に貢献している．NTC は Freq と似た尺度 (順位相関が 0.903) であるので，Freq の代りに精度向上に貢献している場合がある．

これより，個々の組み合せをみると，どの尺度も精度向上に貢献している．しかし，全ての尺度を組み合わせると，どの統合方法においても，精度が最高精度よりも低下している．したがって，組み合わせる尺度は，良く吟味して選択する必要があると言える．

統合の効果および SVM での精度との比較

表4より，F_{mul} による統合の平均精度が最高精度であったので，その統合方法に関して，尺度を追加するごとに，どのような変化が観察できるかを，平均精度，再現率と適合率について見る．また，その性能が，代表的な機械学習の手法である Support Vector Machine (SVM) による方法と比べて，どの程度の精度であるかも見る[7]．

まず，表4などを利用して，1〜5 個の組み合せのうちで，より多くの組み合せが，それより少ないものを包含するようにし，かつ，なるべく平均精度が高くなるようにしたものを選ぶと，表5のようになる．

表5より，Freq に Len を 1 つだけ追加したときの精度の向上が顕著 (8.6 ポイント) であるが，LR_n を加えたときには，1.2 ポイントであり，更に，NTC を加えても精度の向上はわずかで，LR_d を加えると低下していることが分かる．

表 5: F_{mul} による統合の平均精度．

構成尺度	平均精度
Freq, Len, LR_n, NTC, LR_d	0.625
Freq, Len, LR_n, NTC,	0.635
Freq, Len, LR_n	0.633
Freq, Len	0.621
Freq	0.535

これを再現率 (recall) と適合率 (precision) のグラフでみると図5のようになる．なお，図5では，全尺度を統合した場合については，平均精度の低下と呼応し，全体的に適合率が低下するだけであり，グラフが見難くなるので省く．また，図では，複数の構成尺度があるときには，“F_mul:構成尺度” のようにして構成尺度が示されている．また，SVM とあるのは，後述する方法により，SVM を利用した場合の再現率と適合率である．SVM のすぐあとの数字は，利用した構成尺度の数である．

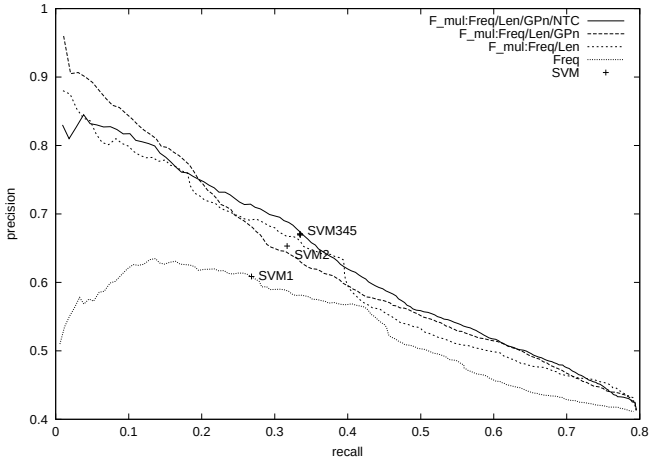

図 5: F_{mul} による統合の再現率と適合率．

図5では，基本となるのは Freq である．これに，Len を加えると，適合率が全体に向上するが，特に，最初の方の，再現率が低い部分での適合率が向上する．この理由は，Freq だけだと，高頻度語が上位にくるのだが，このとき，高頻度の複合名詞は，必ずしも専門用語でないことは良く知られていることであり，そのような複合名詞が上位にくるため，適合率が悪いのだが，Len を追加することにより，ある程度の長さであるような候補が上位にくるようになり，適合率が向上すると考えられる．

次に，LR_n を加えると，更に，立ち上がりの部分の適合率が向上する．これは，連接頻度情報が，良く専門用語の特徴を捉えた尺度であり，かつ，長さや頻度とは異なる情報であるからだと考えられる[8]．

最後に，NTC を加えると，立ち上がりの適合率は低下するが，中盤での精度が向上する．この理由は，それほど明確ではないが，NTC と Freq との順位相関が高いことから，このような頻度に由来する情報を追加することにより，中盤での適合率が向上したのだと考えられる．

次に，SVM を利用して用語候補の各候補について，それが用語かどうかを判定した場合について述べる[9]．このとき，SVM のカーネルとしては線形カーネルを用い，また，素性としては，F_{mul} で利用した尺度についてそれぞれ対数をとったものを用いた．こうすることにより，

$cumLR_d$, LR_n と LR_d の，それぞれについて，0.91, 0.90, 0.91 である．

[7] なお，本稿での問題設定は，1節で述べたように，教師なし学習によるスコアの統合であるので，教師あり学習である SVM との比較は本稿での主旨ではない．しかし，教師なし学習である提案手法と教師あり学習である SVM とを比較することにより，提案手法の性能をより多方面から評価できるため，SVM とも比較することにする．

[8] 順位相関は，$mulLR_n$ と Freq, $mulLR_n$ と Len，Freq と Len のそれぞれについて，0.034, 0.175, -0.213 である．

[9] ソフトウェアは TinySVM(http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/) を利用した．

表 4: 3,4,5 個の組み合せの精度評価.

F_{mul} による統合 (平均精度)	F_{cum} による統合 (平均精度)	Raw_{mul} による統合 (平均精度)
Freq, NTC, Len, LR_n (0.635)	Freq, Len, LR_d, LR_n (0.627)	Freq, NTC, Len, LR_n (0.599)
Freq, NTC, Len, LR_d (0.633)	Freq, Len, LR_d (0.627)	Freq, Len, LR_d (0.596)
Freq, Len, LR_n (0.633)	Freq, Len, LR_n (0.627)	Freq, NTC, Len, LR_d (0.595)
Freq, Len, LR_d (0.630)	Freq, NTC, Len, LR_d, LR_n (0.622)	Freq, NTC, Len, LR_d, LR_n (0.595)
NTC, Len, LR_n (0.627)	Freq, NTC, Len, LR_d (0.622)	Freq, Len, LR_n (0.595)
Freq, NTC, Len, LR_d, LR_n (0.625)	Freq, NTC, Len, LR_n (0.622)	Freq, LR_n (0.594)
NTC, Len, LR_d (0.624)	NTC, Len, LR_d (0.618)	Freq, NTC, LR_d, LR_n (0.588)
Freq, Len, (0.621)	NTC, Len, LR_d, LR_n (0.618)	NTC, Len, LR_n (0.587)
Freq, Len, LR_d, LR_n (0.608)	NTC, Len, LR_n (0.618)	Freq, NTC, LR_n (0.587)
Freq, NTC, LR_d, LR_n (0.606)	Freq LR_n (0.616)	NTC, Len, LR_d, (0.587)
NTC, Len, LR_d, LR_n (0.604)	Freq LR_d, LR_n (0.615)	Freq, NTC, LR_d (0.580)

F_{mul} と同様に各尺度値の乗算をとるだけでなく，更に，SVM により各尺度に重み付けができる．ここで，SVM の訓練およびテストには，全用語候補を利用した．つまり，訓練では，全用語候補に，それが正解候補リストに含まれるか否かをラベル付けし，それを利用してパラメタを調整し，次に，テストでは，同じ全用語候補に対して，正解候補リストに含まれるか否かをテストした．したがって，これは閉じた実験である．このように閉じた実験をした理由は，本稿での比較の対象は，あくまで，教師なし学習における統合法であり，教師あり学習との比較は主眼ではないからである．そのため，ここでは，閉じた実験をし，かつ，カーネル関数も比較のために線形カーネルとした．したがって，このときの適合率や再現率は，F_{mul} で利用した尺度を素性として利用した場合の上限に近い値となると考えられる．

こうした場合の再現率と適合率とを，上述の F_{mul} による尺度の 1〜5 個の組み合せについて，それぞれ，表 6 に示す．また，そのときのグラフ上の点を図 5 に示す．

表 6: SVM での再現率と適合率.

尺度数	再現率	適合率
1	0.268	0.609
2	0.317	0.653
3	0.335	0.671
4	0.335	0.669
5	0.335	0.670

図 5 より，まず，SVM1 が Freq の線上にあるのは，Freq しか素性として利用していないので，当然である．次に，SVM2 をみると，これは，Freq と Len の統合結果より下側にあるので，同じ再現率での適合率は SVM2 の方が低い．SVM345 については，これは，F_{mul} により 4 つの尺度を統合した場合よりも下にあるが，3 つ (と 5 つ) の尺度を統合した場合よりも上にある．また，最高精度という観点からは，SVM345 と 4 つの尺度の F_{mul} による統合とは，同一の再現率での適合率の差は，1 ポイント未満である．

以上より，F_{mul} による統合の結果は，本稿での尺度を利用した場合には，SVM と比べても，遜色のないものであると言える．そのため，F_{mul} による統合は十分効率の良いものであると言える．

5 おわりに

本稿では，複数尺度を統計的に統合する方法を提案し，それらを専門用語抽出という観点から評価した．その結果は，単純に複数尺度の和や積を取る方法と比べると，抽出精度の向上が顕著であった．これより，専門用語抽出においては，提案した統合法が十分に有効であることが分った．今後の課題は，提案手法が有効に機能する場合としない場合およびその理由とを調べることである．

謝辞 筑波大学山本幹雄助教授および通信総合研究所村田真樹主任研究員との草稿段階での議論が参考になった．

参考文献

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*, chap. 3. Addison-Wesley.

Frantzi, K. T. and Ananiadou, S. (1996). “Extracting Nested Collocations.” In *COLING’96*, pp. 41–46.

Frantzi, K. T. and Ananiadou, S. (1999). “The *C-value/NC-value* domain-independent method for multi-word term extraction.” *Journal of Natural Language Processing*, **6** (3), 145–179.

Kageura, K. (1999). “TMREC Task: Overview and Evaluation.” In *Proc. of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, pp. 411–440.

Schone, P. and Jurafsky, D. (2001). “Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem?.” In *EMNLP-2001*, pp. 100–108.

中川裕志, 湯本紘彰, 森辰則 (2003). “出現頻度と連接頻度に基づく専門用語抽出.” 自然言語処理, **10** (1), 27–45.