自然言語処理 122－17
（1997．11．21）

統計量とルールを組み合わせて有用な括弧表現を抽出する手法

久光徹　丹羽芳樹
日立製作所 基礎研究所
〒350-03 埼玉県比企郡鳩山町赤沼2520
{hisamitu, yniwa}@harl.hitachi.co.jp

要旨

新聞記事には平均10行に1個以上の割合で括弧表現，すなわち二つの文字列A，Bが括弧により対応付けられた表現"A（B）"が現れる．このような括弧表現の一部は，“欧州連合（ＥＵ）”や“朝鮮民主主義人民共和国（北朝鮮）”等の言い替えの括弧表現，"日立製作所（会社人事）"等の固有名詞を含む括弧表現であり，これらが特定できれば多数の重要語や固有名詞を獲得できる．本報告では，共起の強さを計る統計指標とエントロピーを字種情報などに基づく単純なルールを組み合わせ，上記の有用な括弧表現を簡便かつ高精度に獲得できることを示す．共起の強さを計る指標として，自己相互情報量，χ^2検定，Yate補正したχ^2検定，頻度，尤度比，Dice係数，改良Dice係数の7種類を比較し，それぞれの効果を調べた．

キーワード：コーパス，情報抽出，統計的手法

Information Extraction from Parenthetical Expressions by Using Statistical Measures and Simple Rules

Toru HISAMITSU and Yoshiki NIWA

Advanced Research Laboratory, Hitachi Ltd.
Hatoyama, Saitama 350-03, Japan
{hisamitu, yniwa}@harl.hitachi.co.jp

Summary

One year worth newspaper articles contain about 300,000 parenthetical expressions. Some of them contain important unregistered words (terms) such as abbreviations, organization names, and company names. The detection of such expressions is therefore an effective way of lexical knowledge acquisition. The proposed method identifies useful parenthetical expressions by combining entropy criteria, a statistical measure to evaluate collocational strength, and a small number of simple rules. In order to select a proper statistical measure, we conducted a comparative evaluation of seven statistical measures: mutual information, χ^2-test, χ^2-test with Yate's correction, frequency, log-likelihood, Dice coefficient, and modified Dice coefficient.

Keywords: corpus-based NLP, information extraction, statistical method

1. はじめに

形態素解析器の高速化に伴い，新聞や特許等の情報検索において形態素解析を用いたインデキシングがしばしば用いられている．この場合の精度低化の要因の一つは形態素解析の解析誤りであり，その原因は主として上記コーパスに頻出する未登録語である．

我々は，未登録人名の獲得についてはすでに報告しており[1][2]，本報告では"EU"，"北朝鮮"のような略称や，社名等の固有名詞の獲得について論じる．

未登録語の獲得に関しては，形態素解析を利用するもの（初期の例として[3]，最近の例として[4]），文字列レベルでの統計的なデータを用いるもの（例えば[5][6]）など，数多くの研究が行われている．

用いるデータの観点からは，例えば形態素解析器にルールベースの未登録語検出機構を組み込み，未登録語を含む局所的な文字列の情報のみから未登録語を獲得するものと，コーパス中に出現する文字や形態素の大域的な統計データを用いて未登録語を獲得するものに分かれる．

略称とその正式名称や会社名を獲得する場合，獲得すべき単位が連語である場合が多いが，前者の枠組みでは連語の認識は容易ではない．一方，後者の枠組みでは，例えばn-gram統計を用いる場合，かなりの計算が必要なだけでなく，取得データのprecisionが低いために結果の選別に多大の労力を要するなどの問題がある．

本報告では，データ抽出の対象領域をうまく選択することにより，統計データに基づく選別と，字種情報などを用いる単純なルールを組み合わせて，上記の情報を簡便かつ高精度に入手できることを示す．

知識獲得の対象領域としては，新聞記事中に出現する「括弧表現」を選んだ．括弧表現とは，"欧州連合（ＥＵ）"や"朝鮮民主主義人民共和国（北朝鮮）"のように，二つの文字列Ａ，Ｂが括弧により対応付けられた表現"Ａ（Ｂ）"のことを指す．括弧表現は新聞記事中に大量に出現し，その一部は複合名詞（組織名を含む）とその省略形など，多くの重要語を含んでいる．そのような括弧表現を選別することにより，大量の言語知識が獲得できる(以下便宜上，Ａを「外側要素」，Ｂを「内側要素」と呼ぶことにする)．

本報告のもう一つの要点は，共起関係の強さを計るために用いる統計的指標の選択に関する．従来，共起関係の強さを計るための指標として，共起頻度，相互情報量(MI)[7]，χ^2等が用いられてきた．中でもMIとχ^2は頻繁に利用されているが，低頻度要素の過大評価を生じるため，頻度の隔たりが大きい対同士を比較する場合は適切でなく，尤度比を用いた方が安定した比較が可能であると指摘されている[8]．これと関連して，これらの指標を統計的手法に基づく漢字文字列の分割に応用した場合の精度を通して，MI，χ^2，尤度比を含む複数の指標の比較を行った研究[9]があり，そこでも尤度比の優位性が示されている．

我々は[10]で本報告と同一のタスクを通して，4種類の統計量の簡単な比較を行ったが，紙面の都合上詳細を述べることはできなかった．本報告では，更にYate補正(Yate's correction)したχ^2検定，Dice係数，改良Dice係数[11]の3種類を加え，7種類の指標の定性的な比較を行った．その結果，我々のタスクにおいては，補正したχ^2検定，改良Dice係数，尤度比の有効性が確認されると同時に，MIとχ^2の問題点が改めて露呈した．なお尤度比に関しては，[8][9]で用いられた表現式を用いず，同等の序列を与えるAIC(またはMDL)の差の形を用いた．

以下，２では，知識獲得対象となる括弧表現を，内容と統計的性質の二つの側面から分析する．３では，統計的基準と簡単なルールの組み合わせによる有用括弧表現の獲得について述べる．共起強度を計る7種類の統計指標の定義と定性的な比較については，４で述べる．

2. 括弧表現の実例と分類

2.1 括弧表現の数

我々が用いた資料（日経新聞1992年1年分）によれば，括弧表現"Ａ（Ｂ）"は，全体で292,799回出現し，その異なり数は177,098であった．このうち2回以上出現したものは，25,421種類であった．

2.2 括弧表現の内容的分類

以下に典型的ないくつかの例をあげる：

(C-I) 言い替え(外側要素を内側要素に置き換え可)
　譲渡性預金（ＣＤ）
　朝鮮民主主義人民共和国（北朝鮮）
(C-II) 読み
　酉（とり）---------- 全体読み
　森本享（すすむ）-------- 部分読み
(C-III) 補足（外側要素と内側要素は交換できない）
　種子島空港（中種子町）
　７月３日（金）
(C-IV) 記事分類・トピック
　日立製作所（会社人事）
　林健太郎著（読書）
(C-V) その他（numbering等）
　理由（１０）

これらの分類は必ずしも網羅的でなく，分類の困難な括弧表現もあるが，多くの有用情報が括弧表現から抽出できることがわかる．特に注目されるのは，(C-I)と(C-IV)の一部の要素である．多くの連語とその省略形が(C-I)型括弧表現の中に，多数の企業名・人名が(C-IV)型括弧表現の中に見いだされる．

2.3 括弧表現における統計的主要型

括弧表現を意味的内容ではなく，外側要素と内側要素の共起頻度だけに注目して統計的な特徴付けをすると，括弧表現には二つの典型的なグループがあることがわかる．すなわち：

(S-I) 1対1型（言い替え型）
　特定の外側要素と内側要素の対が強く共起する．

例）譲渡性預金（ＣＤ）
　　酉（とり）

(S-II) 分類型（多対1型）
特定の内側要素に対して，多数の異なる外側要素が共起するもの．
例）日立製作所（会社人事）

(S-I)型の特徴を持つ括弧表現には，(C-I)や(C-II)型の括弧表現が数多く含まれ，(S-II)型の特徴を持つ括弧表現には，(C-IV)型の括弧表現が多く含まれる．これを知識獲得に生かすことが考えられる．

3. 統計量と言語処理を併用した有用括弧表現の獲得

3.1言い替え型括弧表現の獲得

高頻度(出現回数100以上)の括弧表現だけを眺めれば，殆どが言い替え型の括弧表現であり，例えば次のような簡単なルールのみで有用な表現を選別できるように思われる：

If 外側または内側要素の字種が，英数字または片仮名のみ
　→ 言い換え
Elseif 内側要素の字種が平仮名のみ
　→ 読み
Elseif Ａ，Ｂともに年号
　→ 言い替え
Elseif ＢはＡの弱部分列*
　→ 言い替え（略称）
Else reject

＊ＢがＡの弱部分列とは，Ｂの長さが３以上かつその半分以上が順序を保ってＡに埋め込まれることとする．同様に「ＢがＡの強部分列」とは，Ｂの文字全部が順序を保ってＡに埋め込まれることとする．

しかし，知識獲得対象領域を例えば頻度2以上の括弧表現すべて(異なり数25421)に拡大すると，多くの問題が生じる．例えば，弱部分列条件は，「朝鮮民主主義共和国（北朝鮮）」(頻度942回)のような強部分列条件を満たさない重要表現を獲得するために強部分列条件を緩和したものであるが，この緩和により多数の不適切な括弧表現を獲得してしまう．また，「ＮＫＫ（会社人事）」(頻度11)のように，字種だけでは言い替え型と判定される例も多数出現し，これに対抗して，例えば複数の外(内)側要素に対応する内(外)側要素を持つ括弧表現をrejectすると，表記のゆれや複数の意味を持つ語（譲渡性預金，現金支払機，コンパクトディスクに対応する「ＣＤ」等）が獲得できなくなる．

すなわち，知識獲得の対象となる範囲が広がるにつれ，recallとprecisionを上げるために，ルール毎に条件を緩和したり厳しくしたりする必要が生じ，場合によっては逆に副作用によるrecallとprecisionの低下が起こる．一般に，データ獲得の対象となる集合の品質が悪ければ，ルールの改良だけでは高いデータ獲得精度は望めない．

そこで，ルールの簡易性を保ったまま，知識獲得対象領域側を前処理し，獲得すべき対象を高い割合で，絶対数多く含むようすることを考える．そのための効果的な方法は，統計的な指標を用いて括弧表現を並べ直すことである．

言い替え型の括弧表現"Ａ（Ｂ）"を獲得する場合，用いる統計量はＡ，Ｂの共起強度を計るものを用いることが自然である．共起強度を計る指標としては，MI，χ^2，Yate補正したχ^2検定，頻度，尤度比，Dice係数，改良Dice係数の7種類を比較した結果(これらについては4節で詳述する)，補正χ^2，尤度比，改良Dice係数の3者が，質の良い知識獲得対象領域を得るために有効であるとわかった．

例えば，単純に頻度を用いて，「頻度3以上の括弧表現」を知識獲得対象領域とすると，この中の要素の異なり数は約12,000個であり，その中に含まれる言い替え型括弧の推定数は約3,300個である．一方で，例えば補正χ^2により括弧表現をsortし，最初に頻度1の括弧表現が現れる手前までの括弧表現を知識獲得対象領域として選んだ場合(異なり数6366個)，この中に含まれる言い替え型括弧の推定数も約3,300個であり，同じルール群を用いるならば，この集合を知識獲得対象領域とした方が，効率・精度ともに良い結果が得られることは自明である．

なお，尤度比を用いる場合も，尤度比により括弧表現をsortし，1位から，はじめて頻度1の括弧表現が現れる直前までを判別の対象として選択し(異なり数8,429)，同様の処理を行った．改良Dice係数は頻度1の括弧表現は値が0となるため，要素数がある程度多い知識獲得対象領域としては，頻度2以上の括弧表現全体を取る以外に自然な制限を見いだすことは困難であり，検討中である．

なお，以下に示した実験に適用したルールは，先に述べたルールを若干詳細化したものであり，頻度による順位との比較で，共起が強いと判断された場合の部分文字列条件を緩和し，そうでない場合は緩和しないようにしている．補正χ^2，尤度比，改良Dice係数等は，頻度を反映しつつ，共起強度の弱い括弧表現をかなり効果的に上位から排除するため，以下のルールはこれらの指標を用いた場合かなり有効に働く：

If 外側または内側要素の字種が，
　英数字または片仮名のみ
　　→ 言い換え
Elseif 内側要素の字種が平仮名のみ
　　→ 読み
Elseif Ａを外側要素としたときの内側要素の集合と，Ａを内側要素としたときの外側要素の集合との積集合にBが含まれる
　　→ 言い替え
Elseif Ａ，Ｂともに年号
　　→ 言い替え
Elseif {"頻度によるＡ（Ｂ）の順位" ＞
　　"尤度比によるＡ（Ｂ）の順位"}
　　かつ

"BはAの弱部分列"
→ 言い替え（略称）
Elseif {"頻度によるA（B）の順位" ≦
"尤度比によるA（B）の順位}
かつ
"BはAの強部分列"
→ 言い替え（略称）
Else reject

表1は，各指標で括弧表現をsortした場合，言い替え型括弧がどの程度上位に集まるかを，上位1,000位までの括弧表現を対象に調べたものである．$\chi 2$とMIは，低頻度要素の過大評価のため上位要素の品種が著しく悪い．

表1

	補正χ^2	改良Dice	対数尤度差	頻度
～100位	91(1)	96(1)	90(2)	83(2)
～500位	438(1)	434(13)	418(20)	335(11)
～1000位	807(8)	832(28)	727(46)	554(34)

	χ^2	Dice係数	MI
～100位	16(0)	16(0)	0
～500位	61(1)	61(1)	1(1)
～1000位	114(11)	114(11)	2(1)

*()内の数字は半正解(適切な文字列の削除/追加により正解となる)ものの数

表2に，各指標でsortした上位500個の括弧表現に対する上記ルールの判定精度を挙げる．以下で"正→正"は，目的とする括弧表現を正しく判別した場合，"誤→誤"は，獲得すべきでない括弧表現を正しくrejectした場合で，"正→誤"，"誤→正"（網掛け部分）は誤判別である．

表2

	補正$\chi 2$	改良Dice	対数尤度差	頻度
正→正	438(1)	434(7)	406(16)	308(6)
誤→誤	46	43	72	151
正→誤	14	13	12	27
誤→正	2	10	10	14

	$\chi 2$	Dice係数	MI
正→正	61(1)	61(1)	1(1)
誤→誤	344	344	449
正→誤	0	0	0
誤→正	95	95	50

* ()内の数字は，半正解の数．網掛けは誤り部分

**χ^2とDice係数は，上位に現れる多数の低頻度要素については順序が同じだったため，この表の範囲では区別できない．

また，データ獲得領域内で上位500位以外の部分の知識獲得精度をおおまかに調べるため，200個の括弧表現をデータ獲得領域内から無作為抽出して判別実験を行った結果を表3に示す．Yate補正したχ^2については501位から6366位まで，尤度比については501位から，8,429位までが対象である．改良Dice係数は，便宜上尤度比と併せて501位から，8,429位までで調べた．これらの知識獲得領域内に，3種類のどの基準でも推定約3400個前後の言い替え括弧表現が含まれ，これは頻度2以上の括弧表現全体に含まれる言い替え型括弧表現の推定個数約5100個の2/3を占める．

表3

	補正$\chi 2$	改良Dice	対数尤度差
正→正	113(2)	97(9)	80(9)
誤→誤	72	79	98
正→誤	5	3	3
誤→正	2	21	19

*()内の数字は，半正解の数．網掛けは誤り部分

詳しい定量的な解析は今後の課題であるが，総合的に判断すると，Yate補正したχ^2とルールの組み合わせが効果的である．また，尤度比は，数学的な背景が最も明快であり，精度もさほど悪くなく，普遍的に利用できる指標であるため，χ^2やMIの代わりにbaselineとして利用することが好ましいと思われる．

3.2 分類型括弧表現の獲得

Yate補正したχ^2や尤度比を用いた場合，sortの順序を逆転すれば，分類型(多対1型)括弧表現が上位に現れる．しかし，多対1型括弧表現を獲得することに特化すれば，エントロピーを利用した方が簡単かつ鮮明な結果を得ることができる(逆は成り立たない．すなわち，エントロピーによるsortでは，共起強度をうまく区別できない)．

エントロピーは，次のように用いる．内側要素Bに対して外側要素$A_1,\ldots,A_m$がそれぞれ$f_1,\ldots,f_m$回共起するとき，Bを固定したときのエントロピーE(B)を

$$E(B) = -\sum_{i=1}^{m} \frac{f_i}{F} \log_2 \frac{f_i}{F},\ F = \sum_{j=1}^{m} f_j$$

で定義し，E(B)により内側要素Bをsortする．この結果，例えば"会社人事"，"決算数字"等，数千の企業名と共起する内側要素が上位に現れる．上位語を検討することにより情報抽出上の重要な手掛かりが得られる．表4に，内側要素が数字列でないもの上位10位を示す：

表4

決算数字;11.30	東京; 9.23
会社人事;10.90	本社東京; 9.28
死去;2066;10.82	有価証券含み損; 9.12
業績修正・配当異動; 9.89	仮称; 8.97
ニューフェイス;9.58	読書; 8.90

*内側要素;エントロピー の順．左列が1～5位，右列が6～10位

例えば内部要素として"決算数字"と"会社人事"を選んだとき，異なり数3,255個の企業名が外側要素として獲得された．更に表2中の"本社東京"を一般化した"本社＊"を用いると，異なり数13,205個の企業名が外側要素として獲得された．

4. 共起強度を計る指標

本節では，外側要素Aと内側要素Bの共起の強さを計るための指標を定義する．まずn_{ij} (*i, j = 1, 2*)を次の分割表中で定義する：

	内側要素がB	内側要素が¬B
外側要素がA	n_{11}	n_{12}
外側要素が¬A	n_{21}	n_{22}

これらを用いて，$n_{i.}$とNを次で定義する：

$$n_{i.} = n_{i1} + n_{i2}, n_{.j} = n_{1j} + n_{2j},$$
$$N = \sum_{i,j} n_{ij}.$$

4.1 尤度比

ここでいう「尤度比」とは，注目する内外要素の出現が非独立とした場合と，独立とした場合の最尤推定量の比を指す．尤度比によるsortは，独立最尤モデルと非独立最尤モデルに基づくAIC(又はMDL)の差によるsortと同等であることが簡単に示される．我々は，他の用途も考慮し，[8]の定式化でなく下の形の表現を用いた．ペナルティー項の差から生じる補正項Δは，sortの結果に影響しないが，一般の用途を考慮して付加したままにしてある：

$$2\sum_{i,j} n_{ij}\left\{\log_2\frac{n_{ij}}{N} - \log_2\frac{n_{i.}n_{.j}}{N^2}\right\} + \Delta,$$

AICの場合 ：$\Delta = 2$,

MDLの場合: $\Delta = \log_2 N$.

4.2 Dice係数，改良Dice係数

これらは数学的な背景を持つというより，ヒューリスティックな指標であり，以下で定義される：

$$\text{Dice係数} = \frac{2n_{11}}{n_{1.} + n_{.1}},$$

$$\text{改良Dice係数} = \log_2 n_{11}\frac{2n_{11}}{n_{1.} + n_{.1}}.$$

Dice係数は，低頻度対の比較を相互情報量より精度良くできるとされており，後者はこれを更に改良したものである[11].

4.3 χ^2とそのYate補正

χ^2は以下で定義される：

$$\chi^2 = \sum_{i,j}\frac{\left(n_{ij} - n_{i.}n_{.j}/N\right)^2}{n_{i.}n_{.j}/N} = \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}.$$

χ^2検定は，2項分布$B_i(n,p)$が$np(1-p)>5$を満たすときに正規分布でよく近似されることを基礎にしており，低頻度要素を含む独立性判定には本来適用できない．n_{11}, n_{12}, n_{21}, n_{22}のいずれかが5未満の場合には，Yateの補正と呼ばれる次の式を用いる：

$$\chi^2 = \frac{N(|n_{11}n_{22} - n_{12}n_{21}| - N/2)^2}{n_{1.}n_{2.}n_{.1}n_{.2}}.$$

4.4 MI(相互情報量)

正確には自己相互情報量(self mutual information)と呼ばれ，この場合は次で定義される：

$$MI = \frac{n_{11}/N}{n_{.1}n_{1.}/N^2}.$$

以前から低頻度要素の過大評価が生じることが指摘されているにもかかわらず，しばしば不適切に用いられている．

4.5 各指標について

以上の6種類に頻度を加えた7指標を用い，全括弧表現をsortしたときの上位10位を付録に示した．

MI，Dice係数，χ^2は，A，Bともに相当高頻度の括弧表現に適用を限定しない限り，適切に働かないことが確認された．前2者は，頻度差が反映されないためであり，後者は，既に述べた誤った近似が行われるためである．

補正χ^2，改良Dice係数，尤度比の3者は，基本的には頻度を反映しつつ，共起強度の弱い括弧表現を上位から排除している．特に，補正χ^2を用いた場合，頻度順で第2位の「同（同）」，第7位の「常務（取締役）」，第9位の「退任（取締役）」が，それぞれ42,243位，6,226位，56,796位に後退し，知識獲得対象領域から排除されるか，ごく低順位へと移動させられる．この性質は，簡単なルールによる高精度な知識獲得を助けており，表3の精度に反映している．χ^2の補正については，他の手法も知られており，それらも試みる必要がある．

5. まとめ

新聞記事の括弧表現"A（B）"から有用な情報を抽出するうえで，統計的指標と簡単なルールを組み合わせて略称，固有名詞等を大量に精度良く獲得できる手法を提案した．

我々のタスクにおいては，共起強度を計る指標として，補正χ^2，改良Dice係数，尤度比の3者が有効であった．特にYate補正されたχ^2の望ましい性質があきらかになった．また，共起強度の指標として，MIでなく，尤度比をbaselineとして利用した方が好ましいことを，本稿でも強調しておきたい．

参考文献

[1] 久光徹，丹羽芳樹: 辞書と共起情報を用いた新聞記事からの人名獲得, NL研資料, NL118, pp.1-6 (1997)

[2] Hisamitsu, T., Niwa, Y., and Nitta, Y.: Acquisition of Person Names from Newspaper Articles by Using Lexical Knowledge and Co-occurrence Analysis, Proc. of NLPRS'97 (to appear)

[3] 吉村賢治，武内美津乃，津田健蔵，首藤公昭: 未登録語を含む日本語文の形態素解析，情処論文誌, Vol.30, No.3, pp.294-300(1989)

[4] 朴哲済，筧捷彦:語の連接関係を利用した未知語の形態素辞書情報の獲得手法，自然言語処理, Vol.4, No.1, pp.71-86 (1997)

[5] Nagao,M. and Mori,Y., A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, Proc. of COLING'94, pp.611-615(1994)

[6] 下畑さより，杉尾俊之，永田淳次:隣接文字の分散値を用いた定型表現の自動抽出，NL研資料，NL110-11, pp.71-78 (1995)
[7] Fano, R. (1961). *Transmission of Information.* MIT Press.
[8] Dunning, T., "Accurate Method for the Statistics of Surprise and Coincidence", *Computational Linguistics*, Vol.19, No.1, pp.61-74 (1993)
[9] Kageura, K.: Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences, Internal Memo, Department of Computer Science, University of Sheffield (1996)
[10] 久光徹，丹羽芳樹: 括弧表現から統計量を用いて有用情報を抽出する手法，弟55回情報処理学会全国大会論文集(2)，pp.2-222-2-223 (1997)
[11] 北村美穂子，松本祐治: 対訳コーパスを利用した対訳表現の自動抽出，情処論文誌，Vol.38, No.4, pp.727-735 (1997)

付録
各指標を用いてsortされた括弧表現の上位10位

Yate補正したχ^2

順位	A（B）	頻度
1	マーストリヒト条約（欧州連合条約）	526
2	ロンドン金属取引所（ＬＭＥ）	228
3	米石油協会（ＡＰＩ）	176
4	国際決済銀行（ＢＩＳ）	590
5	国際原子力機関（ＩＡＥＡ）	296
6	欧州通貨制度（ＥＭＳ）	434
7	白金（プラチナ）	394
8	米農務省（ＵＳＤＡ）	62
9	西欧同盟（ＷＥＵ）	55
10	全米自動車労組（ＵＡＷ）	54

改良Dice係数

順位	A（B）	頻度
1	欧州共同体（ＥＣ）	2717
2	独立国家共同体（ＣＩＳ）	1329
3	国連平和維持活動（ＰＫＯ）	1314
4	日本電信電話（ＮＴＴ）	991
5	朝鮮民主主義人民共和国（北朝鮮）	942
6	国際通貨基金（ＩＭＦ）	639
7	国際決済銀行（ＢＩＳ）	590
8	マーストリヒト条約（欧州連合条約）	526
9	国連環境開発会議（地球サミット）	537
10	欧州通貨制度（ＥＭＳ）	434

尤度比（実際はAIC(MDL)の差）

順位	A（B）	頻度
1	欧州共同体（ＥＣ）	2717
2	独立国家共同体（ＣＩＳ）	1329
3	国連平和維持活動（ＰＫＯ）	1314
4	日本電信電話（ＮＴＴ）	991
5	朝鮮民主主義人民共和国（北朝鮮）	942
6	同（同）	1764
7	譲渡性預金（ＣＤ）	764
8	国際通貨基金（ＩＭＦ）	639
9	国際決済銀行（ＢＩＳ）	590
10	マーストリヒト条約（欧州連合条約）	526

頻度

順位	A（B）	頻度
1	欧州共同体（ＥＣ）	2717
2	同（同）	1764
3	独立国家共同体（ＣＩＳ）	1329
4	国連平和維持活動（ＰＫＯ）	1314
5	日本電信電話（ＮＴＴ）	991
6	朝鮮民主主義人民共和国（北朝鮮）	942
7	常務（取締役）	784
8	譲渡性預金（ＣＤ）	764
9	退任（取締役）	703
10	国際通貨基金（ＩＭＦ）	639

χ^2（Dice係数も同じ）

順位	A（B）	頻度
1	種豚（母豚）	1
2	種苗魚（一年魚）	1
3	酒伊エンジニヤリング（福井市）	1
4	酒井謙太郎地区会長（丸万証券会長）	1
5	酒井哲夫教授（養蜂学）	1
6	酒井雄哉大阿闍利（だいあじゃり）	1
7	酒匠アドバイザー（酒販店関係者）	1
8	酒清織物（社長酒井慶治氏）	1
9	酒撰（さけせん）	1
10	酒泉（甘粛省）	1

MI

順位	A（B）	頻度
1	退任（2）	1
2	同（1）	1
3	退任（4）	1
4	同（3）	4
5	日本電信電話（決算数字）	3
6	日本電信電話（会社人事）	5
7	同（下）	1
8	東日本旅客鉄道（決算数字）	1
9	ＪＴＢ（会社人事）	1
10	東海旅客鉄道（会社人事）	1