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

2004－NL－162 (15)
2004／7／16

慣用句抽出のための統計尺度の比較評価

相薗 敏子　小泉 敦子　森本 康嗣

(株)日立製作所 中央研究所

テキストマイニングでは，文書DBから単語や単語のペアを抽出して文書DBの特徴としてユーザに表示する。慣用句は，複数の単語で一つの意味を表すため，全体でまとまった単位として扱う必要がある。本研究では，名詞と動詞のペアからなる慣用句を文書DBから抽出する方式を検討した。本稿では，慣用句を抽出するための統計尺度の比較・評価について述べる。相互情報量，χ^2，AICの比較実験により，χ^2が統計尺度としては最適であることを確認した。また異分野の文書DBを利用することにより，F値が50%から53%に向上した。

Extraction of Japanese noun-verb collocations from Corpora

Toshiko AIZONO　Atsuko KOIZUMI　Yasutsugu MORIMOTO

Central Research Laboratory, Hitachi, Ltd.

Extraction of noun-verb relations is essential for knowledge extraction from text. Proper treatment of collocations is one of the tasks, since they often present non-constitutional concepts and have to be treated as single units instead of verb-noun relations. In this paper, we describe an experimental evaluation of statistical measures for extracting Japanese noun-verb collocations from a corpus. Based on the result of an experiment using three measures; Mutual Information, χ^2, as well as AIC, we conclude that χ^2 is the most suitable measure to extract collocations. F measeure of χ^2 was improved from 50.1% to 52.8% by using a corpus from a different domain.

1. はじめに

企業内における電子化情報の 80%は，報告書やメールのようなテキストデータであるといわれているが，これらには企業にとって価値ある情報が豊富に含まれている。そのため，近年ナレッジマネジメントやCRM(Customer Relationship Management)への期待が高まるなか，これらテキストデータをもっと有効に活用したいというニーズが強まっている。これに応える手段としてテキストマイニングがある。テキストマイニングでは，対象となる分野の文書DBから単語や単語のペアを抽出してユーザに表示し，テキストの分析を支援する。抽出した単語ペアで単語間の相関の強いものには，その分野の特徴を表す表現が多く，文書DBの傾向の把握などには非常に有効である。一方で，それらの中には慣用句が含まれることがある。慣用句は複数の単語で一つの意味を表すため，単語のペアとしてユーザに表示しても利用価値は低い。

慣用句の多くは，名詞と動詞の単語のペアといわれている [1][2]。本研究では，名詞と動詞のペアからなる慣用句を文書DBから抽出する方式を検討した。本稿では，慣用句を抽出するための統計尺度の比較・評価について述べる。

2. 従来研究

慣用句を構成する単語は，出現に相関性が見られる[3]。ある着目した単語に対して相関が強いものを抽出する研究として，相互情報量を用いて着目した単語の類義語を抽出するもの[4]や，χ^2などを用いてテーマを表す名詞と強い相関を持つ要求表現を抽出する研究[5]などがある。

また，ある分野において相関の強い単語ペアの中から慣用句を判別する手法として，シソーラスを用いる研究がある[3]。この研究では，相互情報量を用いて単語間の相関の強さを測っておき，その値が高いペアで，かつ構成単語を類義語で置き換えた単語ペアが存在しない，または存在してもそれらの間の相関の強さともとの単語ペアとの相関の強さに有意な差が認められないとき，慣用句とする。これによれば，相関の強い単語のペアから慣用句の可能性が高いものを絞り込むことが可能となる。

3. 文書DBからの慣用句抽出

文書DBから慣用句を抽出するため本研究では，次のようなアプローチをとる。

(1) 文書DBに出現する名詞と動詞のペアで相関の強いものを抽出し，

(2) 他の分野の文書DBでも出現するものに絞り込んで慣用句の候補とする。

以下詳細に述べる。

3.1 統計尺度を用いた単語ペアの抽出

従来，単語間の相関の強さを測る統計尺度としては，相互情報量，χ^2などが提案されている[4][5]。本稿では，相互情報量，χ^2，AICの3つの統計尺度を実験によって比較し，慣用句の抽出という観点から評価を行う。以下，各統計尺度について説明する。

(1) 相互情報量

相互情報量[3][4]は，ある事象ともうひとつの事象が同時に生起する確率と，それぞれが個別に生起する確率との比に基づいて事象間の結びつきの強さを表す統計量である。単語 W_i,W_j の相互情報量 MI は，次式で求められる。

$$MI(W_i, W_j) = \log_2 \frac{P(W_i, W_j)}{P(W_i)P(W_j)} = \log_2 \frac{m \cdot N^2}{n_i \cdot n_j \cdot M}$$

ただし，n_i ：単語 W_i の出現頻度

n_j ：単語 W_j の出現頻度

N：単語総数

m：単語のペア$\{W_i, W_j\}$の出現頻度

M：単語ペア総数

(2) χ^2

χ^2とは，2組の属性データの統計的検定に用いられる統計量であり，期待値と実際に観測された値のとのずれを表わす[6]。χ^2を用いて単語 W_i, W_j の出現が独立かどうかを検定することで，単語間の相関性を測ることができる。

χ^2の計算では，一般にすべての期待度数が少なくとも5以上であるような標本の大きさが必要とされる。しかし自然言語処理の分野では，全体の事象の数に比して単語が同時に出現する確率は低いため，上記の条件を満たすことは難しい。これに対して本研究では，期待値を修正する「Yatesの補正公式」と呼ばれる次の式を用いる。

$$\chi^2(W_i, W_j) = \frac{M(|ad - bd| - \frac{n}{2})^2}{(a+b)(c+d)(a+c)(b+d)}$$

ただし，a：単語のペア$\{W_i, W_j\}$のペア出現頻度

b：n_i-a　c：n_j-a　d：M-a-b-c

n_i ：単語 W_i の出現頻度

n_j ：単語 W_j の出現頻度　M：単語総数

(3) AIC

AIC(赤池情報量規準:Akaike's Information Criterion)とは，現実に観測されるデータに基づいてモデルの妥当性を検定するための統計量である[7]。単語間の相関性を図るモデルとしては，単語 W_i,W_j の出現が独立であるというモデルと，従属であるというモデルの2つを用いる。各モデルにおいて，事象が生起した度数に基いて最大対数尤度を求め，パラメータの数を用いてAICを計算し，その差 D を単語ペアの関連の強さとして用いる。

$$AIC = AIC_d(W_i, W_j) - AIC_i(W_i, W_j)$$

表1 実験に用いた文書DB

文書DB	利用目的	サイズ	テキスト		文	
		(Kbyte)	数(件)	平均サイズ(byte)	数(文)	平均サイズ(byte)
営業日報DB	慣用句抽出対象DB	22,145.4	58,808	376.57	454,621	48.71
新聞記事DB	絞り込み用異分野DB	33,451.0	41,022	815.44	376,235	88.91

AIC_iは，独立モデルの AIC であり，以下の式で定義される。

$$AIC_i(W_i, W_j) = -2 * ((a+b)\log(a+b) + (a+c)\log(a+c) + (c+d)\log(c+d) + (b+d)\log(b+d) - 2 * M \log M) + 2 * 2$$

AIC_dは，従属モデルの AIC であり，以下の式で定義される。

$$AIC_d(W_i, W_j) = -2 * (a \log a + b \log b + c \log c + d \log d - M \log M) + 2 * 3$$

ただし，n_i :単語 W_iの出現頻度
n_j :単語 W_jの出現頻度
a:単語のペア$\{W_i, W_j\}$の出現頻度
M: 単語ペア総数
b: n_i-a, c: n_j-a, d:M-a-b-c,

以下では，両モデルによるAICの差Dのことを単にAICと呼ぶ。

3．2 異分野DBを用いた単語ペアの絞り込み

相関の強い単語のペアから慣用句の可能性が高いものに絞り込むため，本研究では抽出対象の文書DBとは異なる分野の文書DBを利用する。慣用句は，命題を表す表現に比べ「分野に依存せずに出現する」という特徴があると考えられるからである。例えば，慣用句「難色を示す」と命題的な表現「カタログを渡す」は，ともに構成単語の相関性が強いと考えられるが，後者は営業活動のような分野において特徴的な表現であるのに対して，前者は特定の分野への依存性は比較的低いと考えられる。本研究ではこのような慣用句の特徴に着目し，異なる分野における出現状況に応じて慣用句の候補となる単語ペアを絞り込むこととする。

4．実験および結果の検討

4.1 実験の目的

次の3つの項目について明らかにする。
(1)相互情報量，χ^2，およびAICの比較
(2)慣用句の分野独立性の確認
(3)異分野の文書DBを用いた慣用句候補の絞り込みの効果

統計尺度の比較および異分野の文書DBによる絞り込みの効果を測る指標としてはF値を用いる。

$$\text{F値 } F(\%) = 2 * P * R / (P + R)$$
$$\text{適合率 } P(\%) = \gamma / \beta * 100$$
$$\text{再現率 } R(\%) = \gamma / \alpha * 100$$

ただし，α :文書DBに含まれる慣用句の数，
β :慣用句の候補として抽出された単語ペアの数，
γ :慣用句候補のうち実際に慣用句である単語ペアの数

上記式において「慣用句の候補として抽出された単語ペア」とは，各統計尺度によって単語間の相関が強いと判定されたものを指し，具体的には相関の強い順に上位 100 ペアとする。

評価は主として延べ数ベースで行う。その理由は，抽出対象の文書DBにおいて出現頻度の高い慣用句ほど重要であると考えるからである。

4.2 実験データおよび実験手順

実験には，慣用句の抽出対象の文書DBとして営業日報を格納したDBを用いる。異分野の文書DBとしては毎日新聞経済面に掲載の記事5年分を格納した文書DBを用いる。その詳細を表1に示す。

実験手順を以下に示す。
(1)単語ペアの抽出
営業日報DBおよび新聞記事DBから係り受け関係にある名詞と動詞のペアを抽出する。

表2 営業日報DBにおける正解データ

種別	異なり数		延べ数		例
	数(個)	割合(%)	数(個)	割合(%)	
慣用句	287	4.68	8,671	9.35	{電話,かける},{視野,入れる},{視野,いれる}
慣用句以外	5844	95.32	84,116	90.65	{方針,決める},{デモ,見る},{見積,出す}
合計	6,131	100.00	92,787	100.00	

＊単語ペアの出現頻度が6以上のものを対象とした。

表3.1 統計尺度の比較(異分野DBによる絞り込み前)

統計尺度	延べ数ベース			異なり数ベース		
	適合率(%)	再現率(%)	F値(%)	適合率(%)	再現率(%)	F値(%)
相互情報量	47.4	6.8	11.8	35.0	12.2	18.1
χ^2	58.0	44.1	50.1	43.0	15.0	22.2
AIC	34.6	56.9	43.0	29.0	10.1	15.0

表3.2 統計尺度の比較(異分野DBによる絞り込み後)

統計尺度	延べ数ベース			異なり数ベース		
	適合率(%)	再現率(%)	F値(%)	適合率(%)	再現率(%)	F値(%)
相互情報量	59.4	14.5	23.3	49.0	17.1	25.3
χ^2	52.9	52.6	52.8	47.0	16.4	24.3
AIC	36.2	56.3	44.1	34.0	11.8	17.6

(2)正解データの作成

営業日報DBから抽出した単語ペアに対して慣用句かどうかを人手により判定し正解データを作成する(表2)。

(3)単語ペアの相関の強さの計算

営業日報DBから抽出した単語ペアに対して,相互情報量, χ^2, AICの各尺度を用いて相関の強さを計算する。

(4)絞り込み前の統計尺度の比較

営業日報DBから抽出した単語ペアを相関の強い順に上位 100 位まで取り出し,正解データを用いて適合率,再現率,およびF値を求める(表3.1)。

(5)単語ペアの絞り込み

営業日報DBから抽出した単語ペアのうち新聞記事DBから抽出した単語ペアにも含まれているものを採用,それ以外を削除して,単語ペアを絞り込む。

(6)絞り込み後の統計尺度の比較

絞り込んだ単語ペアを相関の強い順に上位100位まで取り出し,正解データを用いて適合率,再現率,およびF値を求める(表3.2)。

4.3 結果の検討

4.3.1 統計尺度の比較

異分野DBによる絞り込み前(表3.1)と絞り込み後(表3.2)の両方において,延べ数ベースのF値はχ^2が最も高い。また異なり数ベースでみると,延べ数ベースでは次点のAICが最下位となるのに対して,χ^2は最高値,または最高値と同等である。このことからχ^2は,頻度の高いものに偏らず多様な種類の慣用句を抽出できているといえる。よって,慣用句抽出のための統計尺度としてはχ^2が最も適していると考える。

4.3.2 慣用句の分野独立性の確認

慣用句の分野独立性を確認するため,まず営業日報DBから抽出した単語ペアを慣用句とそれ以外に分け,それぞれ新聞記事DBでも出現するものの割合を調査した。その結果を表4に示す。

延べ数において慣用句のうち新聞記事DBでも出現するものの割合(74.5%)は,慣用句以外において新聞記事DBでも出現するものの割合(42.3%)より大きい。すなわち営業分野から抽出した慣用句は,慣用句以外に比べて異分野でも出現しやすいといえる。これを異なり数でみると,慣用句では新聞記事DBでも出現するものの割合(50.9%)と営業日報DBのみで出現するものの割合(49.1%)は同程度だが,慣用句以外では新聞記事DBでも出現するものの割合(28.7%)は延べ数での結果に比べてさらに低い。以上のことから,今回の実験の範囲において慣用句は慣用句

表4 営業日報DBから抽出した単語ペアの異分野DBにおける出現状況

種別	出現状況	延べ数		異なり数		例
		数(個)	割合(%)	数(個)	割合(%)	
慣用句	新聞記事DBでも出現	6,460	74.5	146	50.9	{電話,かける},{視野,入れる}
	営業日報DBのみで出現	2,211	25.5	141	49.1	{データ,落とす},{視野,いれる}
	小計	8,671	100.0	287	100.0	
慣用句以外	新聞記事DBでも出現	35,590	42.3	1,675	28.7	{話,聞く},{ボタン,押す}
	営業日報DBのみで出現	48,526	57.7	4,169	71.3	{名刺,置く},{精度,高める}
	小計	84,116	100.0	5,844	100.0	
合計		92,787	-	6,131	-	

表5 χ^2における上位100ペアの絞り込み前後の比較

種別	絞り込みによる分類		延べ数	異なり数	例
慣用句	Ⅰ	採用	3,402	30	{多岐,わたる},{穴,あける}
	Ⅱ	削除	424	13	{多岐,渡る},{手,離せる}
	Ⅲ	追加採用	1,158	17	{白紙,戻す},{手,つける}
	Ⅰ＋Ⅱ	絞り込み前	3,826	43	
	Ⅰ＋Ⅲ	絞り込み後	4,560	47	
慣用句以外	Ⅰ	採用	1,983	22	{話,聞く},{ボタン,押す}
	Ⅱ	削除	790	35	{名刺,置く},{アライアンス,組む}
	Ⅲ	追加採用	2,073	31	{スピード,速い},{要件,満たす}
	Ⅰ＋Ⅱ	絞り込み前	2,773	57	
	Ⅰ＋Ⅲ	絞り込み後	4,056	53	

以外と比べて分野への依存性が低いことが確認できた。

4.3.3 異分野DBによる絞り込みの効果

異分野DBを用いて単語ペアを絞り込む前(表3.1)と絞り込んだ後(表3.2)の結果を各統計尺度ごとに比較すると，すべての統計尺度においてF値は向上している。ただし，先に慣用句抽出に最も適していると結論したχ^2においては，その向上の割合は 2.7%と小さい。以下，χ^2における絞り込みの効果について詳細に検討する。

表5は，χ^2で上位 100 位以内となった単語ペアを絞り込みの前と後で比較したものである。上位 100 位以内に含まれる単語ペアは，次の3つに分類される。

(Ⅰ)絞り込みにより採用
絞り込み前で上位 100 位以内で，絞り込み後も採用されたもの

(Ⅱ)絞り込みにより削除
絞り込み前では上位 100 位以内であるが，絞り込みによって削除されたもの

(Ⅲ)絞り込みにより追加採用
絞り込み後に上位 100 位以内に入ったもの

上記のうち，タイプⅡで慣用句以外，およびタイプⅢで慣用句に属する単語ペアは，異分野DBを用いた絞り込みの効果に相当する。表5より，前者には「名刺を置く」，「アライアンスを組む」など営業分野の特徴を表す表現が多いことがわかる。

これに対して慣用句でタイプⅡのものは絞り込みの副作用といえる。これらについて詳細に調査したところ，次の3つに分類できることがわかった。

(1)表記・表現のゆれ(5個)
例:「多岐に渡る」
＊新聞記事DBでは「多岐にわたる」のみ出現

(2)営業分野に特徴的な慣用句(4個)
例:「アポをとる」

(3)その他(4個)
例:「手が離せない」

今回の実験で用いた新聞記事DBでは表記・表現が統制されている。そのため(1)のような副作用が生じた。これらは異表記展開などで対応することができる。

(2)のような慣用句は，単語間の相関が強く，かつある分野で特有に出現するという点で，分

野の特徴を表す命題的表現(「カタログを渡す」)と類似している。そのため,相関の強い単語ペアから慣用句の可能性の高いものとして抽出するのは非常に困難である。

また今回絞り込みに用いた文書DBは,抽出対象の文書DBとサイズがほぼ同じである(表1)。(3)に対しては,異分野の文書DBのサイズを大きくするなどにより,IIのタイプから除くことができると考える。

5. まとめ

本研究では,文書DBから統計的なアプローチに基づき慣用句を抽出する方式を検討した。慣用句抽出対象の文書DBとして営業日報DB,異分野の文書DBとして新聞記事DBを用いて実験を行った結果,以下の点を確認した。

(1)単語間の相関の強さを測る統計尺度としてはχ^2が適している。

(2)異分野の文書DBを用いて単語ペアを絞り込むと,F値が2.7%向上した。

今後の課題として,絞り込みに用いる異分野DBに関する検討がある。今回の実験では,営業日報DBに出現する慣用句で新聞記事DBにも出現するものは頻度の高いものが多いことを確認した。その一方で,新聞記事DBを用いて単語ペアを絞り込んでも,χ^2のF値には大きな効果は見られなかった。

今後は,異分野の文書DBとしてサイズや分野が異なるものをいくつか用いて実験を行い,慣用句抽出対象DBと異分野DBとの関係やF値に対する効果などを明らかにする必要がある。

謝辞

新聞記事DBとして,CD-毎日新聞91年版～95年版を毎日新聞社の使用許諾のもとに用いた。

参考文献

[1]奥雅博:日本文解析における述語相当の慣用句の扱い,情報処理学会論文誌,Vol.31,No.12(1990).

[2]首藤公昭,他:日本語の慣用的表現について—語の非標準的用法からのアプローチ—,情報処理学会,自然言語処理研究会報告,No.66-001(1988).

[3]Dekang Lin:Automatic identification of non-compositional phrases, Proc. of the 37th Annual Meeting of the ACL(1999).

[4]前川篤志,伊藤毅志,古郡廷治:係り受け関係と相互情報量を用いた単語の意味獲得,情報処理学会,自然言語処理研究会報告,No.124-008(1998).

[5]山本英子,乾裕子,井佐原均:主観的評価に基づく語幹間関係の評価尺度の比較,言語処理学会第9回年次大会発表論文集(2003).

[6]芝祐順,渡部洋:統計的方法II 推測,社会科学・行動科学のための数学入門3,新曜社(1984).

[7]坂元慶行,石黒真木夫,北川源四郎:情報量統計学,情報科学講座A・5・4,共立出版(1983).