文字、単語統計解析の一手法

1H－6

應江黔　兵藤安昭　池田尚志
岐阜大学工学部

1　はじめに

文字、単語などの自然言語の構成要素の統計解析は、自然言語研究の不可欠な手段である。近年は、単語の頻度、共起頻度などの統計情報をベースに、コンピュータによる単語の自動分類、シソーラスの自動生成などの研究も活発に行なわれるようになった。これらの従来の研究では、テキストデータベース（コーパス）全体にわたって統計され、平均を取った単語の頻度、共起頻度が使われており、単語がコーパスにわたる分布の情報が利用されていない。文字、単語がコーパスにわたる出現頻度の分布関数は、平均、分散、相関などの情報を含んでおり、言語の動的な側面を反映していることから、言語の統計解析に役立つことが期待できる。本稿は単語の出現頻度の分布関数を使って、単語間の相関係数という統計量を導入し、この統計量により単語間の関連性を表すことを検討する。

2　関連性を表す統計量

近年、計算機で処理できる大規模なテキストデータベースが容易に利用できるようになるにつれ、以前は定性的にしか扱えなかった単語の共起、関連関係も定量的に記述できるようになった。単語関連性の尺度として、Churchら[1]が次の相互情報量を提案している。二つの単語x,yについて、それぞれの出現確率をp(x),p(y)とし、両単語が一定の間隔以内に現われる確率をp(x,y)とする。両単語の相互情報量は$I(x,y)=\log_2\frac{p(x,y)}{p(x)p(y)}$と定義される。この他、Smadja[2]は条件確率を使って、固定したある単語に関連の強い単語を抽出する方法を提案している。いずれも同時確率に基づいている。

A statistical method for analyzing word associations,
Jiangqian Ying, Yasuaki Hyodo Takashi Ikeda
Faculty of Enginerring, Gifu University
Gifu-shi,501-11,Japan

ここでは単語の頻度分布から計算される相関係数というもう一つの尺度を提案する。

単語の頻度分布　：テキストデータベースを決まった文数の段落、または決まった単語数の列のようなunitに分割しておく。データベースは順序付きのT個のunitからなるとする。単語xについて、$n_t(x)$を第t unitに現われる回数をとし、$f_t(x)=\frac{n_t(x)}{T}$、$t=1,\ldots,T$とする。$F=(f_1(x),\ldots,f_T(x))$がxの頻度分布となる。

分布の相関係数　：この分布の平均を$\bar{f(x)}$とし、標本標準偏差を$S_x=(1/T\sum_{t=1}^{T}[f_t(x)-\bar{f(x)}]^2)^{1/2}$とする。両単語$x,y$の共分散は$S_{xy}=1/T\sum_{t=1}^{T}[f_t(x)-\bar{f(x)}][f_t(y)-\bar{f(y)}]$、相関係数は$R_{xy}=\frac{S_{xy}}{S_xS_y}$と計算される。

3　実験例

ここでは相関係数を用いて単語間の関連関係を記述することを試みた実例を紹介する。テキストデータベースとして、朝日新聞の１９９１年１月前半の新聞記事約２１万文を用いた。

実験１　３３個の名詞：(農地, 農家, 野菜, 収穫, 職業, 技能, 専門, 育成, 芸術, 作品, 舞台, 演出, 景観, 町並み, 伝統, 電気, 電子, 機械, 化学, 選挙, 立候補, 自民党, 社会党, 文化, 試験, 受験生, 高校生, 大学, 教授, 経済, 会長, 景気, 産業) について、unit=1文のとき、相関係数の大きい２００組を抽出して、更にこれらの単語組の相互情報量を計算した。表１にいくつかの代表的な単語組を示している。この２００組の相関係数の対数を横軸、相互情報量を縦軸にプロットしたのが図１である。そこに両者が比例している傾向が見られる。

実験２　５個の単語（一層、更に、自由自在、臨機応変、柔軟）について、unit=１文、２０文のとき、単語組の相関係数、相互情報量を計算した。結果

をそれぞれ表2，表3に示している。相関係数が負のとき，相互情報量が$-\infty$となっている (unit内に両単語が同時に現れていない)。また，意味的には近いが一文内に共起したことのない単語組 (更に，一層) と (柔軟，臨機応変) がunit=２０文のときに相関係数が大きくなり，相互情報量も正の値になっている。

表 1: 単語組の例 (実験１)

相関係数	相互情報量	単語組
0.793404	8.174316	(機械, 電気)
0.415086	6.766310	(化学, 機械)
0.376831	6.229982	(化学, 電気)
0.319026	7.109216	(機械, 電子)
0.306372	6.628791	(電子, 電気)
0.213180	3.918474	(教授, 大学)
0.209584	5.034002	(受験生, 試験)
0.187378	7.381075	(機械, 技能)
0.179090	4.147733	(大学, 試験)
0.169698	6.841954	(電気, 技能)
0.166231	5.260225	(化学, 電子)
0.152860	6.206376	(町並み, 景観)
0.129804	3.888553	(大学, 受験生)
0.126198	4.175783	(文化, 芸術)
0.116216	5.344193	(機械, 専門)
0.114996	4.211152	(立候補, 選挙)
0.111729	3.420949	(会長, 自民党)
0.104937	4.805072	(電気, 専門)
0.102306	2.848879	(社会党, 自民党)
0.093486	3.788025	(自民党, 立候補)
0.087098	3.276634	(景気, 経済)
0.086626	5.411909	(化学, 技能)
0.085375	4.948812	(産業, 機械)
0.084748	5.079547	(伝統, 町並み)

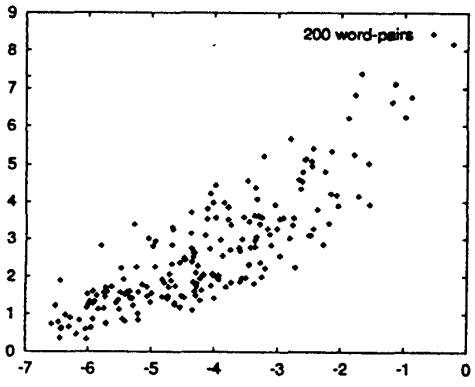

図 1: 相互情報量／相関係数の対数

4 討論

本稿は単語の頻度分布を利用する見地から，単語間の関連性を表す尺度としての相関係数について考察した。初歩的な実験の結果，従来利用されている相互情報量に類似した尺度となっていることを確かめた。これは，言語の分布的な情報の有用性の一端を示している。

表 2: unit=１文の結果 (実験２)

相関係数	相互情報量	単語組
-0.000328	-∞	(更に, 一層)
+0.021304	2.759780	(自由自在, 一層)
-0.000660	-∞	(自由自在, 更に)
-0.000100	-∞	(臨機応変, 一層)
-0.000046	-∞	(臨機応変, 更に)
-0.000202	-∞	(臨機応変, 自由自在)
-0.000441	-∞	(柔軟, 一層)
-0.000202	-∞	(柔軟, 更に)
+0.008434	2.349995	(柔軟, 自由自在)
-0.000062	-∞	(柔軟, 臨機応変)

表 3: unit=２０文の結果 (実験２)

相関係数	相互情報量	単語組
+0.116852	3.167770	(更に, 一層)
+0.095862	1.795405	(自由自在, 一層)
+0.016587	1.047507	(自由自在, 更に)
-0.001756	-∞	(臨機応変, 一層)
-0.000825	-∞	(臨機応変, 更に)
-0.003065	-∞	(臨機応変, 自由自在)
+0.019272	1.273780	(柔軟, 一層)
-0.003515	-∞	(柔軟, 更に)
+0.014334	0.740482	(柔軟, 自由自在)
+0.145841	4.806656	(柔軟, 臨機応変)

従来の統計的言語モデルは，平均の統計データしか使っていない静的なものである。このようなモデルは分野を限定した制限言語には対応できるが，流動的な性質を持つ一般の言語に適用することは難しい。平均データだけでなく，言語の分布的な情報も取り入れ，言語の変動に強い適応型の統計モデルの構築を今後の課題にしたい。

参考文献

[1] K.W.Church and P.Hanks,"Word association norms, mutual information, and lexicography", Computational Linguistics, Vol.16, No.1, 22-29, (1990)

[2] F.Smadja, "Retrieving collocation from Text: Xtract",Computational Linguistics, Vol.19, No.1, 143-177, (1993)