大規模コーパスからの関連語自動抽出

颯々野 学
富士通研究所
sassano@flab.fujitsu.co.jp

1 はじめに

近年、計算機の能力の向上とネットワークの普及により大量の電子化されたテキストが利用できるようになった。我々はこれら大量のテキストから有益な情報や知識を取り出すことを検討している。従来の研究を何を取り出すかという観点から分類すると、(i) 言語外知識を抽出するものと (ii) 言語内知識を獲得するものに大別される[1]。また、どういう手法かという観点から分類すると、(a) 局所的なパターンマッチングを使うものと (b) 統計的な手法を使うものに大別される。(i) に対して手法 (a)、(ii) に対して手法 (b) を利用する研究は盛んに行なわれている (例えば、前者に対しては (Grishman and Sundheim 1996) や、後者に対しては (Church and Hanks 1990; Smadja 1993) など)。

我々のアプローチは (i) に対して手法 (b) を利用するものである。さまざまな種類の大量のテキストを処理する場合には、統計的手法が持つ頑健性が役に立つと考えたからである。ただ、大量のテキストから精度よく詳細な言語外知識を抽出することは難しいので、最初の目標を「大規模なコーパスから統計的な手法を用いて関連語 (連想語や共起語も含む) を抜き出し、その精度や振る舞いを明らかにする」ことに設定した。関連語を抽出するだけでは詳細な言語外知識を取り出したことにはならないが、ある分野の概念を理解するのに類義語や関連語をまとめたシソーラスが役立つこと (Foskett 1980)、関連語のネットワークが情報検索に役立つこと (Doyle 1962) などから考えて十分有益な情報が得られることが期待できる。

本稿では、相互情報量に基づいてコーパスから関連語を自動抽出する方法を提案し、その実験結果を報告する。相互情報量に基づいて言語内知識 (連語) を得る研究は Church と Hanks が行なった (Church and Hanks 1990) が、後に Smadja が指摘しているように Church らの方法は意味的に関連する語のペアを抜き出すものと考えられる (Smadja 1993)。これらの報告をベースにした。

[1] 例えば、言語外知識としては常識や分野固有の知識、5W1H など。一方、言語内知識には文法的知識や語彙的知識などが含まれる (長尾 (編) 1996, pp. 368–371)。

2 関連語抽出の尺度

2.1 基本的な定義

関連語抽出の尺度を与える前に、必要な定義を行なう。以下では単語と形態素を区別せず単語と書く。関連語を抜き出す対象とするコーパスを C と書き、C に出現する単語の集合、C に含まれる全単語数、C に出現する単語異なり数、C に出現する個々の単語、C の先頭から k 番目の単語をそれぞれ、W、N、ω、$w^i (w^i \in W, 1 \leq i \leq \omega)$、$w_k (1 \leq k \leq N)$ とする。

C 中での w^i の頻度を $f(w^i)$ とすると、

$$\sum_{i=1}^{\omega} f(w^i) = N \quad (1)$$

が成り立つ。C 中で w^i が出現する確率 $P(w^i)$ は

$$P(w^i) = \frac{f(w^i)}{N} \quad (2)$$

で定義する。次に、$f_d(w^i, w^j)$ を C 中で w^i と w^j が窓のサイズ d の中に出現する回数とする。ただし、w_k^i と w_l^j は、$1 \leq k < N, 1 < l \leq N, 0 < l - k < d$ を満たすとする。このとき、次式が成り立つ (導出は付録 A 参照)。

$$\sum_{1 \leq i \leq \omega, 1 \leq j \leq \omega} f_d(w^i, w^j) = (d-1)N \quad (3)$$

w^i と w^j がこの順番で窓のサイズ d の中に出現する確率 $P_d(w^i, w^j)$ を次式で定める。

$$P_d(w^i, w^j) = \frac{f_d(w^i, w^j)}{(d-1)N} \quad (4)$$

2.2 順序制約付き単語関連率

事象 A、B が起こる確率をそれぞれ $P(A)$、$P(B)$ とし、事象 A、B が共に起こる確率を $P(A, B)$ とするとき、その相互情報量 $I(A, B)$ は、

$$I(A, B) = \log_2 \frac{P(A, B)}{P(A)P(B)}$$

で表される (Liu 1985, pp. 89–95)。事象 A、B が何らかの従属関係にあり、$P(A, B) > P(A)P(B)$ なら

$I(A,B) > 0$ となる。そこで、A、B をそれぞれ単語 w^i、w^j がコーパス中に出現するという事象に置き換えて考えると、相互情報量は w^i、w^j の関連性を計る指標となることが分かる。

以上の考察に基づき、単語 w^i と w^j との順序制約付き単語関連率を次のように定義する。

$$A(w^i, w^j) = \log_2 \frac{P_d(w^i, w^j)}{P(w^i)P(w^j)} \quad (5)$$

これは窓の扱いを除けば、(Church and Hanks 1990, p. 23) の $I(x,y)$ (association ratio) と同一である。

式 (5) は更に次のように変形できる。

$$A(w^i, w^j) = \log_2 N + \log_2 f_d(w^i, w^j) - \log_2 f(w^i) - \log_2 f(w^j) - \log_2(d-1) \quad (6)$$

2.3 関連度

Church らと異なり、我々の目的は連語の抽出ではないので、語の順序を考慮に入れた $A(w^i, w^j)$ はそのまま使わず、次式で定義される関連度 $r(w^i, w^j)$ を関連語判定の尺度として使う。

$$r(w^i, w^j) = \sqrt{A^2(w^i, w^j) + A^2(w^j, w^i)} \quad (7)$$

ここで $A(w^i, w^j) \geq 0, A(w^j, w^i) \geq 0$ とする。

3 実験および評価

3.1 実験方法

提案手法の有効性を調べるために、日経新聞 1 年分 (日経全文記事データベース日本経済新聞 CD-ROM '96 版) を対象に都市銀行 10 行の関連語を抜き出す実験を行なった。まず、記事の本文を文に分割し、それを Breakfast (颯々野他 1997) で形態素に分割した。この際、特に企業名と地名を辞書に追加している。次に、式 (6)、(7) に従って関連度を計算し、関連語のリスト作成した。処理を簡単にするため、品詞の情報は使わなかった。また、$A(w^i, w^j)$ は低頻度の語の場合に不安定になるので、$f_d(w^i, w^j) \leq 5$ を満たす語については集計しなかった。式 (6) の右辺が負になったときは、$A(w^i, w^j) = 0$ とした。

1996 年の日経新聞 (81,206,223 文字) では、形態素分割済みのデータサイズは 209 MB、形態素の総数 N は 49,536,378 であった。

関連度 r の計算の例を示す。f(さくら銀行) = 2165, f(頭取) = 815, f_{17}(さくら銀行, 頭取) = 64, f_{17}(頭取, さくら銀行) = 8 だったので、A(さくら銀行, 頭取) = $\log_2 49536378 + \log_2 64 - \log_2 2165 - \log_2 815 - \log_2 16 = 6.8$ となる。同様の計算で A(頭取, さくら銀行) = 3.8 を得る。従って r(さくら銀行, 頭取) = $\sqrt{6.8^2 + 3.8^2} = 7.8$ となる。

表 1: 評価結果

	関連語と判定された語の割合 (%)						
r	≥ 4	≥ 5	≥ 6	≥ 7	≥ 8	≥ 9	≥ 10
AS	54	59	56	51	42	58	58
DW	50	56	55	50	46	46	40
FJ	53	56	56	53	49	43	24
HK	53	48	60	76	90	97	97
DC	59	58	55	61	72	77	64
SN	57	54	62	60	54	19	14
SM	55	53	56	47	51	40	32
TK	53	59	59	44	35	87	17
SK	61	64	66	66	64	54	59
TM	60	68	75	88	88	87	86
Av.	56	58	60	60	59	54	49

3.2 評価方法

評価の対象は、1996 年分の記事から抜き出した都市銀行 10 行の関連語とし、それらを人手で評価した。なお、窓のサイズは 3、5、9、17、33、65 の中から実験的に 17 を選んだ。関連度 $r(w^i, w^j)$ 4 以上を評価対象にした。評価対象のデータを表 2 に示す (スペースの都合上 r が 5 以上のもののみ載せた)。

実際には 21 人の評価者に、抜き出されたデータが 1996 年の日経新聞から自動抽出したものだと告げ、関連語とみなせないと判断されるものをチェックしてもらった。

3.3 評価結果

関連度 r の閾値を 4 から 10 まで変えたときの評価結果を表 1 に示す。AS などのアルファベット 2 字は、個々の銀行を表す略号である (表 2 参照)。以下では、関連語であると判定されたもの割合を適合率とする。最下行の銀行 10 行平均を見ると、閾値 5 から 8 で適合率はおよそ 60 % である。しかし、この数値の解釈には注意が必要である。評価者が持っている知識が反映されているからである。固有名詞に関する関連語の関連性を厳密に判定しようとすれば、その語に関する知識が要求される。

そこで、評価者のうち一定の割合 (m % とする) 以上の人が関連語と認めたものは何らかの関連があるとみなし、そのときの適合率を計算した。北海道拓殖銀

行 ($r \geq 6$) の例 (表 2 参照) で説明する。抽出された関連語 11 語それぞれについて何人の人が関連語と判定したか調べる。そして、評価者 21 人中 $m = 10$ % (2 人) 以上の人が関連語と判定したものに印を付ける。仮に 8 語に印がついたとするとき、適合率を $8/11 = 73$ % と計算するのである。この計算を銀行 10 行 ($r \geq 6$) について行ない、平均を取ると 85 % となった。同様に $m = 20, 30$ とすると、適合率はそれぞれ 76 %、68 % となった。関連性を広くとらえたい場合にはこちらの指標のほうが参考になる。

上記の二つの定量的な評価以外に、実際に抜き出された語を見てみると非常に興味深い。$r \geq 10$ にはその銀行固有の事柄 (頭取の名前や、合併前の銀行名など) がよく並び (表 1 で閾値が高いところで適合率が下がるのは、人名などを関連語でないと判定した人が多かったため)、$r = 6$ 辺りには同業の他の銀行名などが上がっている。更に下には金融関連の言葉などが見られる。今回の実験で品詞情報は使わなかったが、助詞や活用語尾などは現れていない。単純に共起確率が高いものから関連語を抜き出す方法では、これらを取り除く必要があるが、本手法では不要である。

また、銀行ごとの違いを見ても、それぞれ特徴的な違いが出ている。例えば、大和銀行では巨額損失事件、富士銀行ではテニス (沢松奈生子選手の所属は富士銀行) の話題が出ている。

4 おわりに

本稿では、相互情報量を基づいてコーパスから関連語を抽出する方法を提案し、新聞記事 1 年分から都市銀行の関連語の抽出実験の結果を報告した。今後の課題として他の関連語抽出法との比較がある。相関ルール抽出などデータマイニングで知られた手法 (例えば (Agrawal et al. 1996) など) をテキストデータに適用してみることや、相互情報量以外の言語知識獲得の手法を適用してみることが考えられる。今回の実験では低頻度のものを ad hoc に削除したが、適切な検定を行なって閾値を設定することも必要である。

謝辞 評価に協力してくれた富士通研究所メディア統合研究部の方々に感謝します。

参考文献

Agrawal, Rakesh; Mannila, Heikki; Srikant, Ramakrishnan; Toivonen, Hannu; and Verkamo, A. Inkeri (1996). "Fast Discovery of Association Rules," In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, MIT Press, 307–328.

Church, Kenneth Ward and Hanks, Patrick (1990). "Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, **16**(1), 22–29.

Doyle, Lauren B. (1962). "Indexing and Abstracting by Association. Part I," System Development Corporation Research Paper, SP-718/001/00.

Foskett, D.J. (1980). "Thesaurus," In A. Kent, H. Lancour, and J.E. Daily (Eds.), *Encyclopedia of Library and Information Science*, **30**, 416–462.

Grishman, Ralph and Sundheim, Beth (1996). "Message Understanding Conference - 6: A Brief History," In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, 466–471.

Liu, Chung Laung (1985). "Elements of Discrete Mathematics, Second Edition," McGraw-Hill.

長尾 真 (編) (1996). 自然言語処理, 岩波ソフトウェア科学, **15**, 岩波書店.

颯々野 学, 斎藤 由香梨, 松井 くにお (1997). アプリケーションのための日本語形態素解析システム, 言語処理学会第 3 回年次大会発表論文集, 441–444.

Smadja, Frank (1993). "Retrieving Collocations from Text: Xtract," *Computational Linguistics*, **19**(1), 143–177.

A 式 (3) の導出

ある位置 $l(1 \leq l \leq N)$ にある単語 w_l に対して、$\max(0, l-d) < k < l$ を満たす単語 w_k の個数 $c_d(l)$ は次式で与えられる。

$$c_d(l) = \begin{cases} l-1 & (1 \leq l \leq d) \\ (d-1) & (d < l \leq N) \end{cases} \tag{8}$$

$c_d(l)$ は、$max(0, l-d) < k < l$ を満たす単語対 (w_k, w_l) の個数を与えている。$f_d(w^i, w^j)$ の総数は $c_d(l)$ を 1 から N まで足し合わせたものに等しい。

$$\sum_{1 \leq i \leq \omega, 1 \leq j \leq \omega} f_d(w^i, w^j) = \sum_{l=1}^{N} c_d(l) \tag{9}$$

更に $d \ll N$ の場合は次のように近似できる。

$$\sum_{l=1}^{N} c_d(l) = (d-1)N \tag{10}$$

式 (9) と式 (10) から式 (3) を得る。

表 2: 都市銀行 10 行の抽出結果 (関連度 5 以上)

あさひ銀行 (AS) 協和銀行 (13.4), 埼玉銀行 (12.3), 芳樹 (10.7), 大坂 (10.1), 敦 (8.4), 浜田 (8.1), あさ (8.0), 調査部長 (8.0), 為替 (7.4), さくら銀行 (7.1), 住友銀行 (6.8), 地名 (6.7), 興銀 (6.5), 信託銀行 (6.2), 支店長 (5.7), 次長 (5.7), 頭取 (5.7), 現 (5.1)
大和銀行 (DW) 安部川 (10.3), 巨額 (10.2), 重罪 (9.9), 海保 (9.8), 据置 (9.4), 住友銀行 (9.3), 津田 (9.1), ベアリングズ (9.0), 司法 (8.9), 井口 (8.9), 秀幸 (8.8), 澄夫 (8.8), リミテッド (8.4), 検察 (8.3), 損失 (8.1), ニューヨーク (8.1), 連邦 (7.9), 信託 (7.9), 隠匿 (7.8), 仲 (7.8), 撤退 (7.4), 発覚 (7.4), 起訴 (7.3), 事件 (7.1), 不正 (7.1), 配当率 (7.0), 行員 (6.8), 大和 (6.7), 頭取 (6.6), 無罪 (6.6), 有罪 (6.6), 裁判 (6.6), 金銭 (6.5), プレミアム (6.4), 不祥事 (6.4), 十一億 (6.3), 反論 (6.2), 罪 (6.1), 支店 (5.9), 罰金 (5.7), 当局 (5.7), 証言 (5.6), 2月 (5.6), 格付け (5.5), 地検 (5.5), 住友商事 (5.5), さくら銀行 (5.4), 7月 (5.4), 取引 (5.3), 地裁 (5.2), 大蔵省 (5.2), ヒット (5.2), 富士銀行 (5.1), 住専問 (5.1), 地名 (5.0), 契機 (5.0)
富士銀行 (FJ) 沢松奈 (11.8), 境米夫 (11.7), 生子 (11.6), 芳春 (10.5), 綾部 (10.3), 第一勧業銀行 (9.1), 全国銀行協会連合会 (9.0), 東京三菱銀行 (8.8), 住友銀行 (8.2), 徹 (8.1), さくら銀行 (7.9), 頭取 (7.9), 収 (7.8), 日本興業銀行 (7.7), オペル (7.4), シングルス (7.3), 為替 (7.2), 入行 (6.9), 回戦 (6.8), 北海道拓殖銀行 (6.5), 三和銀行 (6.3), 東海銀行 (6.3), 都銀 (6.2), づら (6.0), 富士 (5.9), 営業部 (5.8), 地名 (5.7), シード (5.7), 次長 (5.5), 橋本 (5.4), 伊達 (5.3), ローン (5.2), 杉山 (5.2), 資金 (5.1), 大和銀行 (5.1), バンク (5.0), 固定 (5.0)
北海道拓殖銀行 (HK) 北海道銀行 (11.6), 北洋銀行 (11.0), 札幌銀行 (10.7), 優先株 (8.2), 武蔵 (7.9), 第一勧業銀行 (7.1), 1日 (6.6), 富士銀行 (6.5), 0 (6.4), 地名 (6.4), 札幌市 (6.3), 道 (5.8), 不良 (5.8), 8月 (5.5), 人事 (5.5), 閉鎖 (5.4), 9 7 (5.4), 1月 (5.3),. (5.0)
第一勧業銀行 (DK) 角倉雅 (12.0), 第一銀行 (10.6), 東京三菱銀行 (9.1), 富士銀行 (9.1), 正司 (8.3), さくら銀行 (8.2), 三和銀行 (8.2), 宝くじ (7.8), 浩一 (7.5), 池田 (7.4), 為替 (7.3), 裕 (7.2), 北海道拓殖銀行 (7.1), 小切手 (7.0), 克己 (7.0), 中田 (6.7), 次長 (6.7), 一郎 (6.5), 地名 (6.5), 東海銀行 (6.3), 住友銀行 (6.1), 奥田 (6.0), 資金 (5.9), 証券 (5.9), メーン (5.6), 地合い (5.5), 頭取 (5.4), 日本興業銀行 (5.4), 国際 (5.3), 偽造 (5.3), 固定 (5.3), 支店長 (5.1), 部 (5.1)
三和銀行 (SN) 堅二 (11.9), 上席 (10.3), 川勝 (9.8), 満 (8.8), 住友銀行 (8.6), 東京三菱銀行 (8.6), さくら銀行 (8.5), 第一勧業銀行 (8.2), 日本興業銀行 (7.9), 入行 (7.5), 為替 (7.1), 斎藤 (7.0), 大田区 (6.9), ドル買い (6.7), 営業部 (6.5), 富士銀行 (6.3), 地名 (5.9), 低利 (5.6), 役 (5.6), 木津 (5.6), 雪 (5.5), 売り (5.2), 頭取 (5.2), 谷 (5.0), 代理 (5.0)
住友銀行 (SM) 広行 (14.7), 野手 (13.4), 弘一 (12.5), 巽外夫 (11.8), 住友クレジットサービス (10.2), 三木 (9.5), 大和銀行 (9.3), 俊明 (8.9), 三和銀行 (8.6), さくら銀行 (8.6), 富士銀行 (8.2), 河合 (7.6), 入行 (7.2), 敏雄 (7.1), 2 0 0 1 (6.9), あさひ銀行 (6.8), ファイナンス (6.5), 東京三菱銀行 (6.3), 東海銀行 (6.2), 第一勧業銀行 (6.1), 代理 (6.1), 振り込み (5.9), 頭取 (5.7), メーン (5.6), 次長 (5.5), 1月 (5.5), 1日 (5.4), ナショナル (5.4), 地合い (5.1), 人事 (5.1), 4月 (5.1)
東海銀行 (TK) 喜一郎 (13.2), 西垣 (12.6), 夏樹 (10.2), 道標 (8.8), 信用組合 (8.4), さくら銀行 (8.3), 覚 (8.2), 優先株 (7.9), 譲渡 (7.3), 中部電力 (6.8), 頭取 (6.8), 日本長期信用銀行 (6.5), 調査部 (6.5), 調査部長 (6.4), 富士銀行 (6.3), 高木 (6.3), 第一勧業銀行 (6.3), 住友銀行 (6.2), 破たん (6.1), ロビー (5.8), 相談役 (5.6), 信組 (5.5), 検証 (5.5), 詐欺 (5.3), 東海 (5.3), 1月 (5.0), ローン (5.0)
さくら銀行 (SK) 太陽神戸銀行 (11.1), 俊作 (10.7), ムーディーズ (10.7), 鳴海 (10.3), 三井銀行 (10.1), さくら (10.0), A 2 (9.9), 全国銀行協会連合会 (9.6), A 3 (9.4), ミディアム (9.3), 保証付き (9.2), ターム (9.2), ファイナンス (8.7), 住友銀行 (8.6), 三和銀行 (8.5), 東海銀行 (8.3), 第一勧業銀行 (8.2), 富士銀行 (7.9), 劣後 (7.8), 頭取 (7.8), 営業部 (7.7), 優先株 (7.7), 太平洋銀行 (7.7), ルクセンブルク (7.4), あさひ銀行 (7.1), 全銀協 (6.9), 足立 (6.8), 格付け (6.7), 主任 (6.7), 調査部 (6.6), 借り換え (6.5), 都市銀行 (6.5), 彰 (6.3), 為替 (6.2), 日本興業銀行 (5.8),〉 (5.7), 東京三菱銀行 (5.7), ローン (5.6), 〈(5.5), 大和銀行 (5.4), 参考人 (5.4), 証券 (5.4), 郵便 (5.4), 地名 (5.3), 都銀 (5.2), 役 (5.2), 資金 (5.1), 橋本 (5.1), メーン (5.1), 変更 (5.0)
東京三菱銀行 (TM) 商工組合中央金庫 (12.2), 三菱銀行 (11.6), 東京銀行 (11.2), 富士 (10.0), 第一勧業銀行 (9.1), 富士銀行 (8.8), 日本信託銀行 (8.7), 三和銀行 (8.6), 日本興業銀行 (8.3), 為替 (7.9), 利付 (7.8), 金融債 (7.6), − (6.8), 農林中央金庫 (6.4), 日本債券信用銀行 (6.4), 調査部 (6.4), 康 (6.4), 住友銀行 (6.3), 地名 (6.1), 日本長期信用銀行 (6.0), 誠 (5.8), 証券 (5.7), さくら銀行 (5.7), 宮崎 (5.7), 資金 (5.7), 誕生 (5.5), 売り出 (5.5), 合併 (5.2), 小林 (5.1), 次長 (5.0)