自然言語処理 139－9
（2000．9．22）

線形結合モデルを用いたトピック分析

李 航　山西 健司

NEC 情報通信メディア研究本部

〒216-8555 川崎市宮前区宮崎 4-1-1

{lihang,yamanisi}@ccm.cl.nec.co.jp

本稿では，テキストの「トピック構造」を解明することを目的とする新しいテキスト処理，トピック分析を提案する．トピック構造はテキストがどのようなトピックからなり，トピックがテキスト内でどのように変化するかを表すものである．また，本稿では，統計的学習手法によるトピック分析を提案する．具体的には，トピックを単語のクラスタによって表現し，トピック分析に線形結合モデルを用いる．実験結果によれば，本研究の方法が従来法の組み合わせより分析精度が著しくよいことがわかった．

Topic Analysis Using A Finite Mixture Model

Hang LI　Kenji YAMANISHI

C&C Media Research, NEC Corporation

4-1-1 Miyazaki Miyamae-ku Kawasaki, 216-8555 Japan

Abstract

We address the issue of 'topic analysis,' by which is determined a text's topic structure, which indicates what topics are included in a text, and how topics change within the text. We propose a novel approach to this issue, one based on statistical modeling and learning. We use word clusters to represent topics, and employ a finite mixture model to represent a word distribution within a text. Our experimental results indicate that our method significantly outperforms a method that combines existing techniques.

1 はじめに

本稿では新しいテキスト処理であるトピック分析を提案する．ここでいうトピック分析とは，テキストがどのようなトピックからなり，それらがテキスト内でどのように変化するかといったテキストのトピック構造を解明することである．トピック分析は主にトピック同定とテキスト分割という２つのタスクからなると考えられる．

トピック分析は幾つかの応用において極めて有用である．例えば，テキストに対してトピック分析を行うことによって，テキストがどのようなトピックからなり，それぞれのトピックがテキスト内のどの部分で言及されているかについて判断することができる．このようなトピック分析の結果を索引にし，情報検索に利用することができる．

従来では，上記の意味でのトピック分析の研究がなされていなかった．関連研究としてテキスト分割とキーワード抽出に関する研究がそれぞれあった．しかし，提案されたテキスト分割法 (例えば，(Hea97)) は，テキストをトピックの変化に応じて幾つかのブロックに分割できるが，分割された各ブロックにおけるトピックの同定を行えなかった．一方，提案されたキーワード抽出法 (例えば，(SY73)) は，テキストからトピックを表すキーワードを抽出できるが，テキスト分割を行えなかった．

本研究では，統一した枠組でトピック分析，つまりテキスト分割とトピック同定を行うことを目指す．

本研究の方法の主な特徴は，A) 単語のクラスタ (単語の集合) を用いてトピックを表現し，B) 線形結合モデルを用いてテキスト内の単語の分布を表現することである．線形結合モデルは二層の確率分布の階層構造をもつ．第一層はトピックの確率分布からなる．第二層はトピック内の単語の確率分布からなる．単語の確率分布はさらにトピックの確率分布によって線形結合される．以下このような線形結合モデルを確率的トピックモデル (Stochastic Topic Model，或はSTM) と呼ぶ．

トピック分析を行う前に，コーパスにある単語の共起データを基に単語のクラスタ，即ちトピックを作成する．特に，我々はMDL原理 (Ris96) を用いた単語クラスタリングの新しい方法を開発した．

トピック分析では，テキストの一つのブロックが一つのSTMによって生成されると仮定して，テキストを生成した可能性のもっとも高いSTMを推定する．実際，STM間の距離の変化の大きさに基づいてテキストをブロックに分割し，分割された各ブロックにおけるSTMを推定し，確率値の高いトピックを同定する．このように得られる，複数のブロックとそれぞれのブロックのトピックをテキストのトピック構造とする．

別々に提案された従来法を組み合わせることによってトピック分析を実現することができる．具体的には，例えば，従来のキーワード抽出法 (SY73) を用いてテキストからキーワードを抽出し，抽出されたキーワードをトピックとみなし，従来のテキスト分割法 (例えば，(Hea97)) を用いてテキストを分割し，分割されたブロックにおけるトピックの分布を推定し，確率値の高いトピックを同定することによってトピック分析を行うことができる．我々の実験結果によれば，本研究の方法はこのような従来法の組み合わせより精度が著しくよいことがわかった．例えば，トピックの同定では本研究の方法は従来法の組み合わせより再現率 (recall) が11%ほど上回る (0.515 vs 0.405) ことがわかった．本研究の方法が高い精度を達成できるのは，単語クラスタを用いることと確率モデルを用いることによるところが大きい．

線形結合モデルは従来ではテキスト分類，情報検索等に用いられていたが (例えば，(LY97; Hof99))，トピック分析に用いられるのは本研究が初めてであると思われる．また，本研究における線形結合モデルの定義の仕方と利用の仕方は従来研究のそれとは異る．

2 確率的トピックモデル

2.1 トピック

トピックという概念は言語理論によって定義が異る．本研究ではそれをテキスト内に現われる話し項目として捉える．特にトピックを話し項目と関連する単語のクラスタによって表現する．また，単語クラスタには中心となるシード単語が存在すると仮定する．図1が示すのは単語「trade」をシードとする単語クラス，即ちトピックの例である．

trade: export import tariff trader GATT protectionist

図 1: トピックの例

2.2 モデル

W が単語の集合であるとし，K がトピックの集合であるとする．まず，トピックの確率分布 $P(k) : \sum_{k \in K} P(k) = 1$ を定義する．次に，それぞれのトピック $k \in K$ に対して，単語の確率分布 $P(w|k) : \sum_{w \in W} P(w|k) = 1$ を定義する．但し，w が k に含まれない時，$P(w|k)$ が0であるとする．次に，単語の確率分布 $P(w|k)$ をトピックの分布 $P(k)$ によって線形結合したモデルを確率的トピックモデル (Stochastic Topic Model，或はSTM) と定義する．よって，単語 w の確率は

$$P(w) = \sum_{k \in K} P(k)P(w|k) \quad w \in W.$$

と計算される．以下，$P(w|k)$ と $P(k)$ をパラメータとよぶ．

本研究では，テキストを複数のSTMによって生成される単語の列とみなす．しかも，一つのSTMがテキストの一つのブロック (或は，パラグラフ) を生成すると仮定する．これらのSTMは同じトピックの集合からなるが，異るパラメータをもつとする．図2は線形結合モデルの例を示す．

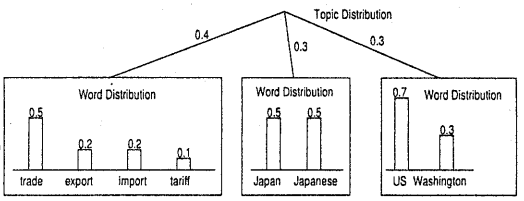

図 2: 確率的トピックモデルの例

3 単語クラスタリング

トピック分析を行う前に，コーパスデータを基に単語のクラスタ，すなわちトピックを作成する．具体的には，

すべての単語をシード単語とし，それぞれのシード単語に対して，それとコーパスデータにおいて共起しやすい単語を集め，一つの単語クラスタにする．例えば，図1に示すのは単語「trade」をシードとする単語クラスタである．

我々は確率的コンプレキシティ，即ちMDL原理(Ris96)を用いて共起単語を集める新しい方法を開発した．

いま，シード単語 s とその他の任意の単語 w の共起の度合を計算するとする．データは m 個のテキストであるとする．s と w に関わるデータは $(s_1, w_1), (s_2, w_2), \cdots, (s_m, w_m)$ となる．但し，(s_i, w_i) は単語 s と単語 w の i 番目のテキストにおける出現状況を表す．$s_i \in \{1, 0\}$, $w_i \in \{1, 0\}$, $(i = 1, \cdots, m)$, 1は単語が出現したことを表し，0は単語が出現しなかったことを表す．また，$s^m = s_1 \cdots s_m$ と $w^m = w_1 \cdots w_m$ と表す．

まず，単語 s に依存しない，単語 w の生成モデル I(即ち，ベルヌイモデル)を考える．w^m のモデル I に対する確率的コンプレキシティ(SC)を計算する：

$$SC(w^m : I) = mH\left(\frac{m^+}{m}\right) + \frac{1}{2}\log\frac{m}{2\pi} + \log\pi,$$

ここで，m^+ は w^m の中の1の数を表し，log は底が2である対数を表し，π は円周率を表す．また，$H(z) \stackrel{\text{def}}{=} -z\log z - (1-z)\log(1-z)$, when $0 < z < 1$; $H(z) \stackrel{\text{def}}{=} 0$, when $z = 0$ or $z = 1$ とする．

一方，単語 s に依存する，単語 w の生成モデル D を考え，w^m のモデル D に対する確率的コンプレキシティ(SC)を計算する：

$$SC(w^m : D) = \left(m_s H\left(\frac{m_s^+}{m_s}\right) + \frac{1}{2}\log\frac{m_s}{2\pi} + \log\pi\right) + \left(m_{\neg s} H\left(\frac{m_{\neg s}^+}{m_{\neg s}}\right) + \frac{1}{2}\log\frac{m_{\neg s}}{2\pi} + \log\pi\right)$$

ここで，m_s は s^m における1の数を表し，$m_{\neg s}$ は s^m における0の数を表し，m_s^+ は w^{m_s} における1の数を表し，$w_{\neg s}^+$ は $w^{m_{\neg s}}$ における1の数を表す．但し，w^{m_s} は対応する s_i が1である w_i $(w_i \in w^m)$ の部分列を表し，$w^{m_{\neg s}}$ は対応する s_i が0である w_i $(w_i \in w^m)$ の部分列を表す．

次に，

$$\delta SC = \frac{1}{m}\Big(SC(w^m : I) - SC(w^m : D)\Big) = \left[H\left(\frac{m^+}{m}\right) - \frac{m_w}{m}H\left(\frac{m_w^+}{m_w}\right) - \frac{m_{\neg w}}{m}H\left(\frac{m_{\neg w}^+}{m_{\neg w}}\right)\right] - \left\{\frac{1}{2m}\log\frac{m_w m_{\neg w}\pi}{2m}\right\}. \quad (1)$$

を計算する．MDL原理[1]によれば，δSC の値が大きければ，w の出現状況が s の出現状況により依存する．[2]

実際，δSC の値が与えられた閾値 γ 以上かつ $P(w|s) > P(w)$ が成り立つ単語 w をシード単語 s と共起しやすい単語とする．

単語クラスタリングはトピック分析とは独立したもので，ここで他のクラスタリング法(例えば，(Hof99))を用いることも可能である．通常の単語クラスタリング法(例えば，(Hof99))の処理時間のオーダーは $O(|D||W|^2)$ であるが，本研究のは $O(|D| + |W|^2)$ のみである．但し，$|D|$ はクラスタリング用テキストの数を表し，$|W|$ は単語の種類の数を表す．従って，テキストの数が多い時，本研究のクラスタリング法がより実用的である．

また，本研究の単語クラスタリング法は通常の単語クラスタリング法にない利点をもっていることにも注目されたい．同義語(例えば，「Britain」と「UK」)がほとんど同じテキストに現われないので，通常のクラスタリング法でそれらを同じクラスタにまとめることが困難である．しかし，本研究の方法の場合，同義語が関連単語(例えば，「London」)をシードとする同じクラスタにまとまる可能性が高い．

4 トピック分析

4.1 入出力

本研究のトピック分析では，与えられたテキストを解析し，そのテキストのトピック構造，即ちテキストにおける複数のブロックとそれぞれのブロックのトピックを出力する．図3はテキストのトピック構造の例を示す．トピック構造から，テキストがどのようなトピックからなり，トピックがテキスト内でどのように変化するかをみることができる．

トピック分析では，テキストのメイントピックとサブトピックも出力できる．例えば，図3ではシード単語「trade-export-tarrif-import」によって表現されるものはメイントピックで，「Japan-Japanese」や「Hong-Kong」等によって表現されるものはサブトピックである．

4.2 処理の流れ

トピック分析は，前処理としてのトピック発見，テキスト分割とトピック同定の三つの過程からなる．トピック発見では，対象テキスト内に現われるトピックを収集する．収集されたトピックを基にSTMを作成することができる．テキスト分割では，一つのSTMが一つのブロックを生成していると仮定し，STM間の距離の変化の大きさを基に対象テキストを分割する．トピック同定では，分割された各ブロックのSTMを推定し，確率値の高いトピックを同定する．このように得られる，複数のブロックとそれぞれのブロックのトピックを対象テキストのトピック構造とする．

4.3 トピック発見

トピック発見では，まず対象テキストからキーワードを抽出する．具体的には，対象テキスト内の各単語のシャノン情報量と呼ぶ量を計算する．単語 w のシャノン情報量は

$$I(w) = -N(w)\log P(w)$$

と定義される．ここでは，$N(w)$ は w の対象テキスト内の出現頻度で，$P(w)$ はコーパスデータから推定された w の出現確率である．$I(w)$ は単語 w によって表現される「情報の量」であると解釈することができる．実際，I の値の降順にソートされた上位 l 個の単語を対象テキストのキーワードとする．

シャノン情報量は情報検索等で広く使われているtf-idf(SY73)という量に近い．情報理論の立場からみれば，シャノン情報量の利用は正当化できるが，tf-idfの利用は正当化が困難である．本研究の予備実験では，シャノン情報量のキーワード抽出の精度がtf-idfのに較べて同等以上であることもわかった．

[1] MDL原理の解説は(Li98)を参照されたい．

[2] 式(1)における $[\cdots]$ 部分は相互情報量(cf.,(BPd+92))であることに注目されたい．式における $\{\cdots\}$ 部分は，データサイズが小さい時相互情報量が大きくなることを抑える役割をもつ．

ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RIFT (25-MAR-1987)

block 0 ——— trade-export-tariff-import(0.12) Japan-Japanese(0.07) US(0.06)
0 Mounting trade friction between the U.S. and Japan has raised fears among many of Asia's exporting nations that the row could inflict ...
1 They told Reuter correspondents in Asian capitals a U.S. move against Japan might boost protectionist sentiment in the U.S. and lead to ...
2 But some exporters said that while the conflict would hurt them in the long-run, in the short-term Tokyo's loss might be their gain.
3 The U.S. Has said it will impose 300 mln dlrs of tariffs on imports of Japanese electronics goods on April 17, in retaliation for Japan's ...
4 Unofficial Japanese estimates put the impact of the tariffs at I0 billion dlrs and spokesmen for major electronics firms said they would ...
5 "We wouldn't be able to do business," said a spokesman for leading Japanese electronics firm Matsushita Electric Industrial Co Ltd &lt.
6 "If the tariffs remain in place for any length of time beyond a few months it will mean the complete erosion of exports (of goods subject ...

block 1 ——— trade-export-tariff-import(0.17) US(0.09) Taiwan(0.05) dlrs(0.05)
7 In Taiwan, businessmen and officials are also worried.
8 "We are aware of the seriousness of the U.S. threat against Japan because it serves as a warning to us," said a senior Taiwanese trade ...
9 Taiwan had a trade trade surplus of 15.6 billion dlrs last year, 95 pct of it with the U.S.
10 The surplus helped swell Taiwan's foreign exchange reserves to 53 billion dlrs, among the world's largest.
11 "We must quickly open our markets, remove trade barriers and cut import tariffs to allow imports of U.S. products, if we want to defuse ...
12 A senior official of South Korea's trade promotion association said the trade dispute between the U.S. and Japan might also lead to ...
13 Last year South Korea had a trade surplus of 7.1 billion dlrs with the U.S., up from 4.9 billion dlrs in 1985.
14 In Malaysia, trade officers and businessmen said tough curbs against Japan might allow hard-hit producers of semiconductors in third ...

block 2 ——— Hong-Kong(0.16) trade-export-tariff-import(0.10) US(0.05)
15 In Hong Kong, where newspapers have alleged Japan has been selling below-cost semiconductors, some electronics manufacturers share ...
16 "That is a very short-term view," said Lawrence Mills, director-general of the Federation of Hong Kong Industry.
17 "If the whole purpose is to prevent imports, one day it will be extended to other sources. Much more serious for Hong Kong is the ...
18 The U.S. last year was Hong Kong's biggest export market, accounting for over 30 pct of domestically produced exports.

block 3 ——— trade-export-tariff-import(0.14) Button(0.08) Japan-Japanese(0.07)
19 The Australian government is awaiting the outcome of trade talks between the U.S. and Japan with interest and concern, Industry ...
20 "This kind of deterioration in trade relations between two countries which are major trading partners of ours is a very ...
21 He said Australia's concerns centred on coal and beef, Australia's two largest exports to Japan and also significant U.S. ...
22 Meanwhile U.S.-Japanese diplomatic manoeuvres to solve the trade stand-off continue.

block 4 ——— Japan-Japanese(0.12) measure(0.06) trade-export-tariff-import(0.05)
23 Japan's ruling Liberal Democratic Party yesterday outlined a package of economic measures to boost the Japanese economy.
24 The measures proposed include a large supplementary budget and record public works spending in the first half of the financial year.
25 They also call for stepped-up spending as an emergency measure to stimulate the economy despite Prime Minister Yasuhiro Nakasone ...
26 Deputy U.S. Trade Representative Michael Smith and Makoto Kuroda, Japan's deputy minister of International Trade and Industry (MITI),...

0-26: 文 id
block 0-4: ブロック id
例: Hong-Kong(0.16) trade(0.10)
トピック「Hong-Kong」と「trade」がそれぞれ 0.16 と 0.10 の確率で出現

図 3: テキストのトピック構造

$k_1, \ldots, k_n$ はクラスタ，$V = \{\{k_i\}, i = 1, 2, \cdots, n\}$
各 (k_i, k_j) に対して，k_i のシードが k_j に含まれ，
k_j のシードが k_i に含まれれば，(k_i, k_j) を Q に入れる
while $(Q \neq \emptyset)$ {
　Q から最初の要素 (k_i, k_j) を取り出す
　if (k_i と k_j が異る集合 W_1 と W_2 に属する)
　　W_1 と W_2 を $W_1 \cup W_2$ で置き換える
}
V の各 W に対して，W の要素を一つのクラスタにマージ

図 4: マージング・アルゴリズム

トピック発見では，次に，予め作成された単語クラスタを参照し，抽出されたキーワードをシードとする単語クラスタを全部集める．

次に，もし二つのクラスタのシードが互いのクラスタ内に含まれるのであれば，その二つのクラスタをマージする．例えば，tradeをシードとする単語クラスタが単語importを含み，importをシードとする単語クラスタが単語tradeを含む場合，この二つの単語クラスタをマージする (図4を参照)．このようなマージ処理を経て，図3のような複数の単語「trade-import-tariff-export」をシードとする単語クラスタが得られる．

このように，対象テキスト内におけるもっとも顕著かつ互いに独立なトピック (クラスタ) が収集される．

4.4 テキスト分割

テキスト分割では，まず分割の候補となる点を決める．説明の便宜上ここでは比較的に短いテキストの分割を考える．その場合，分割候補点は各文の文末となる．それぞれの分割候補点に対して二つの疑似テキストを作成する．その一つは候補点より前の h 個の文からなり，もう一つは後の h 個の文からなる (文の数は h より少ない場合は存在する文だけとする)．各候補点に対して，前後の疑似テキストを基に，EM アルゴリズム ((DLR77), 図5を参照) を用いて前後の STM をそれぞれ推定する．

$l = 1, \cdots, t$ と繰り返す (t は自然数)

$$P^{(l+1)}(k|w) = \frac{P^{(l)}(k)P^{(l)}(w|k)}{\sum_{k \in K} P^{(l)}(k)P^{(l)}(w|k)}$$

$$P^{(l+1)}(k) = \frac{N(w)P^{(l+1)}(k|w)}{N}$$

$$P^{(l+1)}(w|k) = \frac{N(w)P^{(l+1)}(k|w)}{\sum_{w \in W} N(w)P^{(l+1)}(k|w)}$$

$N(w)$ は w の出現頻度，$N = \sum_{w \in W} N(w)$

図 5: EM アルゴリズム

その後，前後の STM の類似度[3](距離の逆) を計算する．具体的には，前後の STM が $P_L(w)$ と $P_R(w)$ である時，類似度は

$$S(L||R) = 1 - \frac{\sum_{w \in W} |P_L(w) - P_R(w)|}{2}$$

[3] 類似度を用いるのは従来法との比較を容易にするためである．

```
n は分割候補点の数, θ は閾値
各候補点の類似度 S(i)(i = 0 ... n) を計算
for (i = 1; i < n − 1; i++){
    if (S(i − 1) > S(i) & S(i + 1) > S(i)){
        j = i − 1;
        while (j > 0 & S(j − 1) > S(j))
            j − −;
        P1 = S(j);
        j = i + 1;
        while(j < n & S(j + 1) > S(j))
            j + +;
        P2 = S(j);
        if (P1 − S(i) > θ & P2 − S(i) > θ)
            i でテキストを分割
}}
```

図 6: 分割アルゴリズム

と計算する．分子の部分は統計学における **variational distance** で，二つの確率分布の距離を測る尺度としてよい性質をもつ ((CT91), p.299)．

このように対象テキストの各候補点に対して類似度を計算し，各候補点の類似度のグラフ (図 4.4) を作成することができる．グラフにおける「谷」は分割すべき点となる．実際，ある谷の類似度が左と右の山の類似度との差が与えられた閾値 θ 以上である場合，その谷の点で対象テキストを分割する (図 6を参照)．例えば，図 3の例では，$\theta = 0.05$ の場合，番号 6，14，18，と 22 の文の文末でテキストを分割する．

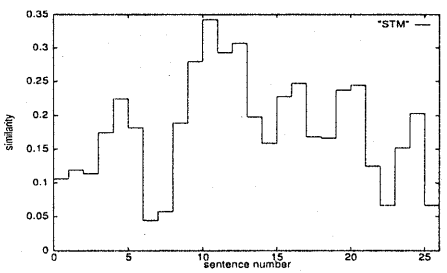

図 7: 各候補点の類似度

4.5 トピック同定

テキスト分割の後，各ブロックの STM のパラメータを EM アルゴリズムを用いて推定し，各ブロックのクラスタ (トピック) の確率分布を得る．次に，各ブロックにおいて確率値の高いクラスタ (トピック) を選び，図 3で示されるようなトピック構造を構築する．ここでは，トピックはシード単語によって表現される．

また，全ブロックを通じて選ばれるトピックをメイントピックとし，一部のブロックにおいてだけ選ばれるトピックをサブトピックとする．

5 応用

トピック分析は幾つかの応用に利用することができる．

例えば，与えられた一連のテキスト (例：ホームページ) に対してトピック分析を行い，それらのテキストに現われるトピックの索引を作成することができる．どのトピックがどのテキストのどのブロックに言及されているかを示し，さらに，どのトピックがメイントピックで，どのトピックがサブトピックであるかを示すことができる．ユーザがこの索引をみることによってそれらのテキストの内容を大まかに把握することができる．また，ある特定のテキストに対して，そのテキストのトピック構造を示すことができ，ユーザがそれをみることによってそのテキストの内容をおよそ知ることができる．

トピック分析は，この他に，テキストマイニング，テキスト要約，情報抽出等のテキスト処理にも利用可能である．というのは，これらの応用では，まずテキストの構造を明らかにする必要があるからである．

6 関連研究

我々の知っている限りでは，従来では同じ枠組内でトピック分析，つまりテキスト分割とトピック同定を行う方法が提案されていなかった．

広く使われている従来のキーワード抽出法 (SY73) では，テキスト内に現われるすべての単語の tf-idf の値を計算し，tf-idf の大きい単語をそのテキストのキーワードとする．このように抽出されたキーワードをトピックとみなすことができる．しかし，従来のキーワード抽出法はテキストの分割を行うことができない．

従来では数多くのテキスト分割法が提案されていた．しかし，従来のテキスト分割法はトピック同定を行うことができない．

従来のテキスト分割法は二つのグループに分けることができる．その一つは，一つのテキストをそのトピックの変化に応じて幾つかのブロックに分割するもので，もう一つは，つながっている幾つかのテキストを元のテキストに分割するもの (例えば，(ACD+98; YCG+98; BBL99; Rey99)) である．前者は通常教師なし学習に基づき，後者は通常教師あり学習に基づく．特に，Hearst の TextTiling の方法 (Hea97) は前者のグループの代表的なものであり，本研究の設定ともっとも近い．TextTiling では，単語の頻度ベクトルの間の余弦を類似度とし，類似度の大きさの基にテキストの分割を行う．

一方，本研究の方法は，同じ枠組内でテキスト分割とトピック分析を行うことを特徴とする．確かに，従来法を組み合わせることによってトピック分析を実現することができる．具体的には，tf-idf を用いてテキストからキーワードを抽出し，抽出されたキーワードをトピックとみなし，TextTiling を用いてテキストを分割し，分割されたブロックにおけるトピックの分布を推定し，各ブロックにおいて確率値の高いトピックを同定することによってトピック分析を行うことができる．しかし，我々の実験では，本研究の方法は従来法の組み合わせより精度が著しくよいことがわかった．これは，本研究の方法が単語クラスタを用いることと確率的モデルを用いることによるところが大きいと思われる．もっとも重要なのは，本研究の方法が図 3に示すようなトピック構造を出力できるが，従来ではトピック構造という考えがなかった点である．

トピック分析はテキスト分類 (LSCP96; LY97; Joa98; WAD+99; NMTM00)) とは異る処理であることに注意されない．テキスト分類ではカテゴリを用いるが，トピック分析ではカテゴリの概念がない．テキスト分類の出力はカテゴリを表すラベルであるが，トピック分析の出力

はテキストのトピック構造である．また，通常テキスト分類は教師あり学習によるものが多いが，トピック分析は教師なし学習によるものである．

線形結合モデルは幾つかのテキスト処理に使われている (例えば，(LY97; NMTM00; Hof99))．しかし，従来の線形結合モデルの定義の仕方と利用の仕方は本研究のそれとは異る．例えば，Li と Yamanishi が線形結合モデルを用いたテキスト分類法を提案した (LY97)．線形結合モデルはカテゴリの上で定義されている：

$$P(w|c) = \sum_{k \in K} P(k|c)P(w|k), w \in W, c \in C,$$

ここでは，W は単語の集合，C はカテゴリの集合を表す．Hofmann が情報検索に線形結合モデルを用いることを提案している (Hof99)．具体的には，「アスペクトモデル」と呼ばれるモデルをつかう．

$$\begin{aligned} P(w,d) &= P(d)P(w|d) \\ &= P(d)\sum_{k \in K} P(k|d)P(w|k) \\ & w \in W, d \in D \end{aligned}$$

ここでは，D はテキストの集合を表す．アスペクトモデルは同時分布のモデルである．

7 実験結果

本研究の方法と従来法の組み合わせによるトピック分析の比較実験を行った．比較は，得られるトピック構造の妥当性，テキスト分割の精度，トピック同定の精度の点で行われた．

7.1 データ

トピック分析のベンチマークデータがないため，実験ではテキスト分類でよく使われているロイター通信データ[4]を用いた．同データにある 9603 のトレーニングテキストを単語クラスタリングに，3299 のテストテキストをトピック分析に用いた．

各テキストに対して Oxford Learner's Dictionary[5]を用いてその単語を原型へ変換した．また，各テキストから我々用意した不要語辞書を用いて不要語を削除した．このように得られた一テキストの平均長さは 115 単語であった．

7.2 実験手順

9603 のトレーニングテキストを用いて単語クラスタを作成した．5 回以上の出現頻度をもつ単語が 7340 あった．これらの単語をシード単語とし，単語のクラスタを作成した．クラスタリングの閾値 γ は 0.005 に設定した．シードを含めて 2 つ以上の単語をもつ単語クラスタが 970 あった．トレーニングテキストが 90 のカテゴリに分類され，カテゴリを表すラベルがつけられているが，単語クラスタリングにはラベルが必要でないので，それを使わなかった．

次に，3299 のテストテキストに対してトピック分析を行った．トピック分析の閾値とパラメータ θ，l，h は予備実験結果に基づきそれぞれ 0.05，20，3 に設定した．

[4] http://www.research.att.com/ lewis/ から入手可能．
[5] ftp://sable.ox.ac.uk から入手可能．

7.3 定性評価

本研究の方法による 3299 のテストテキストに対するトピック分析の結果が人間の直観と一致するかどうかを評価した．

この実験のトピック分析では，各ブロックにおいてクラスタをその確率値の降順でソートし，ソートされたクラスタのシードを上位 7 つ選び，そのブロックのトピックを表すとした．

図 3が示すのは分析結果の一例である．図 3の例では，テキストが 5 つのブロックに分割され，各ブロックが上位 7 つのシードに表現される．また，メイントピックはシード「trade-export-tariff-import」によって表現され，サブトピックは例えばシード「Japan-Japanese」によって表現されている．得られたトピック構造は人間の直観とほぼ一致していることがわかる．しかし，幾つの誤りもあった．テキストは，文の内容が少ないため文 11，13 の文末で分割されなかった．また，シャノン情報量の値が高いことによって「(Mr.) Button」が誤ってトピックとして選ばれた．

7.4 メイントピック同定

次に，3299 のテストテキストのトピック構造からメイントピックを同定し，これらのメイントピックとテストテキストに振られているカテゴリを表すラベルとおよそマッチするかどうかの評価を行った．

カテゴリを表すラベルは学習 (単語クラスタリング) には用いていなく，テストだけに用いていることに注目されたい．これはラベルが学習とテストの両方に用いられる通常のテキスト分類の状況とは異る．従って，ここで得られるメイントピック同定の結果は通常のテキスト分類の結果と比較することができない．

本研究の方法 (これを STM とよぶ) では，各ブロックにおいてクラスタをその確率値の降順でソートし，ソートされたクラスタのシードを上位 k 個選び，そのブロックのトピックを表すとする．また，全ブロックを通じて選ばれるトピック (実際，シード単語) をメイントピックとする．テキストが分割されない場合，つまり，一つのブロックしかない場合，そのブロックに選ばれるすべてのトピックをメイントピックとみなす．

表 1では，もっとも大きい 10 のカテゴリとそれぞれのカテゴリに属すテストテキストの数を示す．我々は，それぞれのカテゴリに対して，その定義に従って，認定語とよぶ単語を定義した．表 1では，各カテゴリの認定語を示す．

STM に対して，メイントピックと判定されたシード単語が認定語を一つでも含む場合，STM がその認定語のカテゴリ (つまり，メイントピック) を正しく同定できたとする．

評価では，「再現率」と「適合率」を用いた．再現率は正しく同定できたテキストの同定すべきテキストの中の占める割合で，適合率は正しく同定できたテキストの同定したテキストの中の占める割合である．

6 節で述べた従来法の組み合わせ (これを Com とよぶ) によるトピック分析も行った．Com の場合，テキストを分割し，分割された各ブロックにおいてキーワードをその確率値でソートし，ソートされた上位 k 個のキーワードをそのブロックのトピックとした．さらに，すべてのブロックに選ばれたトピック (キーワード) をメイントピックとした．メイントピックと認定されたキーワードが認定語を一つでも含む場合，Com がその認定語のカテゴリ (つまり，メイントピック) を正しく同定できたとする．

表 1: カテゴリと認定語

カテゴリ	テキスト数	認定語
earn	1087	earning, share, profit, dividend
acq	719	acquisition, acquire, sell, buy, merger
money-fx	179	currency, dollar, yen, stg
grain	149	grain, cereal, crop
crude	189	oil, crude, gas
trade	117	trade, export, import, tariff
interest	131	interest & rate
ship	89	ship, vessel, ferry, tanker, shipping
wheat	71	wheat
corn	56	corn, maize

表 2: メイントピック同定結果 (トップ 7 単語)

カテゴリ	STM		Com	
	再現率	適合率	再現率	適合率
earn	0.790	0.971	0.526	0.976
acq	0.245	0.854	0.184	0.841
money-fx	0.436	0.456	0.285	0.421
grain	0.322	0.750	0.174	0.650
crude	0.487	0.676	0.407	0.664
trade	0.667	0.473	0.590	0.356
interest	0.107	0.700	0.084	0.733
ship	0.247	0.957	0.270	0.828
wheat	0.620	0.936	0.408	0.967
corn	0.429	0.960	0.446	1.00
micro-average	**0.515**	**0.824**	0.365	0.774

表 3: メイントピック同定結果 (トップ 5 単語)

カテゴリ	STM		Com	
	再現率	適合率	再現率	適合率
earn	0.742	0.971	0.348	0.977
acq	0.184	0.868	0.120	0.869
money-fx	0.413	0.503	0.268	0.471
grain	0.295	0.759	0.121	0.600
crude	0.471	0.718	0.333	0.656
trade	0.479	0.505	0.513	0.403
interest	0.053	0.700	0.069	0.818
ship	0.169	1.000	0.180	0.762
wheat	0.577	0.953	0.282	0.952
corn	0.357	0.952	0.321	1.000
micro-average	**0.461**	**0.850**	0.257	0.767

表 2と 3では，k が 5 と 7 の時の STM と Com のメイントピック同定の結果を示す[6]．STM と Com が同じ数の単語でメイントピックを表現する場合の評価なので，公正な評価と言える．結果から，本研究の方法は従来法の組み合わせより再現率と適合率ともに大きく上回ることがわかった．

これは本研究の方法が単語ではなく，単語のクラスタを用いることに起因する．例えば，単語「trade」，「import」，「export」，「tarrif」があるブロックにそれぞれ一回ずつ出現したとする．このブロックでは，貿易がトピックになっていると推測することができる．しかし，Com が貿易のトピックを同定するのが困難である．というのは，上記単語のそれぞれの出現頻度が低いからである．一方，STM が上記単語を一つのクラスタにまとめているため，そのクラスタ全体の出現頻度が高く，貿易のトピックを同定できる可能性も高い．

7.5 テキスト分割

任意に選んだ二つのテキストをマージすることを繰り返し，500 の疑似テキストを作成した．本研究の方法 STM と従来法 Com を用いてこれらの疑似テキストを分割し，分割が正しいかかどうかを評価した．評価には再現率と適合率の他に，誤り確率を用いた．誤り確率とは任意に選んだ k(この場，$k = 5$) 文離れた二つの文が誤って分割された確率のことである (BBL99)．

表 4は分割の結果を示す．結果から本研究の方法が従来法よりいずれの評価尺度においても分割精度がよいことがわかった．考察によれば，これは本研究の用いる距離尺度 variational distance によるところが大きいことがわかった．

[6]micro-average の定義については (LR94) を参照．

8 おわりに

本稿では，確率的トピックモデルと呼ばれる線形結合モデルを用いたトピック分析を提案した．

トピック分析は，主にテキスト分割とトピック同定からなる．トピック分析によってテキストのトピックの構造を解明することができる．

本研究の方法は以下の特徴をもつ．1) 単語のクラスタでトピックを表現し，確率的トピックモデルでテキスト内の単語分布を表現する．2) トピック分析を行う前にコーパスデータから単語クラスタを作成する．3) STM 間の距離の変化の大きさを基にテキスト分割を行う．4) STM のパラメータを推定することによってトピック同定を行う．

我々の実験結果によれば，本研究の方法は従来法の組み合わせよりテキスト分割の精度やトピック同定の精度が良いこともわかった．

参考文献

J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot

表 4: テキスト分割結果

カテゴリ	STM			Com		
	再現率	適合率	誤り率	再現率	適合率	誤り率
earn	0.660	0.660	0.167	0.640	0.640	0.171
acq	0.820	0.820	0.059	0.740	0.740	0.085
money-fx	0.700	0.700	0.087	0.660	0.660	0.121
grain	0.700	0.700	0.074	0.660	0.660	0.076
crude	0.860	0.860	0.051	0.820	0.820	0.066
trade	0.800	0.800	0.072	0.800	0.800	0.081
interest	0.760	0.760	0.119	0.820	0.820	0.084
ship	0.837	0.854	0.074	0.816	0.833	0.084
wheat	0.760	0.760	0.075	0.640	0.640	0.130
corn	0.625	0.625	0.147	0.650	0.650	0.105
平均	**0.752**	**0.754**	**0.092**	0.725	0.726	0.100

study: Final report. *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machi. Lrn.*, 34:177–210, 1999.

P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. Class-based n-gram models of natural language. *Comp. Ling.*, 18(4):283–298, 1992.

T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons Inc., New York, 1991.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journ. of Roy. Stat. Soci., Ser. B*, 39(1):1–38, 1977.

M. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Comp. Ling.*, 23(1):33–64, 1997.

Thomas Hofmann. Probabilistic latent semantic indexing. *Proc. of SIGIR'99*, pages 50–57, 1999.

T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Proc. of ECML'98*, 1998.

H. Li. *A Probabilistic Approach to Lexical Semantic Knowledge Acquisition and Structural Disambiguation.* Ph.D. Thesis, Univ. of Tokyo, 1998.

D. D. Lewis and M. Ringuette. A comparison of two learning algorithms for test categorization. *Proc. of 3rd Ann. Symp. on Doc. Ana. and Info. Retr.*, pages 81–93, 1994.

D. D. Lewis, R. E. Schapire, J. P. Callan, and R. Papka. Training algorithms for linear text classifiers. *Proc. of SIGIR'96*, 1996.

H. Li and K. Yamanishi. Document classification using a finite mixture model. *Proc. of ACL'97*, pages 39–47, 1997.

K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Machi. Lrn.*, 39:103–134, 2000.

J. C. Reynar. Statistical models for topic segmentation. *Proc. of ACL'99*, pages 357–364, 1999.

J. Rissanen. Fisher information and stochastic complexity. *IEEE Trans. on Info. Thry.*, 42(1):40–47, 1996.

G. Salton and C.S. Yang. On the specification of term values in automatic indexing. *Journ. of Doc.*, 29(4):351–372, 1973.

S. M. Weiss, C. Apte, F. Damerau, F. J. Oles, T. Goetz, and T. Hampp. Maximizing text-mining performance. *IEEE Intel. Sys.*, 14(4):63–69, 1999.

J.P. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. A Hidden Markov Model approach to text segmentation and event tracking. *Proc. of ICASSP'99*, pages 333–336, 1998.