自然言語処理 108－1
(1995. 7. 20)

確率モデルによる主題の自動抽出

野本 忠司
(株) 日立製作所 基礎研究所
〒 350-03 埼玉県比企郡鳩山町赤沼 2520
email: nomoto@harl.hitachi.co.jp

あらまし
本稿では日本語テキストにおける主題の自動抽出に向けて、新しい手法を提案する。本稿では、主題抽出を一種の文書分類 (Text Categorization) と捉え、従来の文書分類の技術を応用した確率的な主題抽出方式を提案する。また、本稿では、格助詞を利用し主題抽出における文法情報の効果について検討する。

ＣＤ－ＲＯＭ版日本経済新聞 (1992 年 1 ～ 6 月) の 42,401 件の記事をデータとして利用し、 格情報あり、格情報なしの条件下で主題抽出の実験を行った。また、評価の基準としてベースラインを導入した。実験の結果では、格情報を利用したモデルが他の場合に比べ優位であることが確認された。しかし、本方式は記事が長くなるにつれ精度が急速に低下するなどの問題が見られ、改善の余地があることも分った。

和文キーワード　　日本語談話、テキスト分類、主題抽出

A Probabilistic Approach to Topic Identification

Tadashi Nomoto
Advanced Research Laboratory, Hitachi Ltd.
2520 Hatoyama Saitama 350-03 Japan

Abstract
The paper describes a new method for discovering topical words in discourse. It shows that text categorization techniques can be turned into an effective tool for dealing with the topic discovery problem. Experiments were done on a large Japanese newspaper corpus. It was found that training the model on annotated corpora does lead to an improvement on the topic recognition task.

英文 key words　　**Japanese Discourse, Text Categorization, Topic Identification**

1 はじめに

主題 (topic) の決定問題は、談話研究の中でも重要な課題であるが、この問題は言語学あるいは工学的な立場ら様々な検討がなされてきた。(機能主義) 言語学における主題とは、文中にあって先行文脈に現れた周知の事柄や物 (旧情報) を表し、一方、対立概念である**焦点** (focus) は構造的、形態的あるいは音韻的に強調された語句であり、先行文脈にはない新しい情報を表すと言われる (Kuno, 1973)。主題が照応現象 (co-reference) に関わるのはこのような事情からである。

自然言語処理では、主題と焦点について厳密な区別をしないのが一般的であるが、言語学と同様に主題の問題は代名詞の解釈との関連で議論されてきた。つまり、代名詞の解釈を決定している文脈中の要素 (名詞句) が主題であるという立場である (Sidner, 1983; Grosz and Sidner, 1986; Walker et al., 1994)。一方、情報検索の分野ではテキストを予め用意しておいた項目 (カテゴリー) に分類するという考え方がある (Lewis, 1992; Finch, 1994; Iwayama and Tokunaga, 1994)。項目はそれに分類されたテキストの一種の主題 (話題) と受け取れるが、分類項目とテキストは独立に与えられるため、言語学で言う「主題」と意味合いが異なる。

本稿では、一応主題と照応問題を切り離して、テキスト分類の立場からテキストの特徴語 (salient word) としての主題の抽出を考えることにする。ただし、テキスト内容を分類項目に、より直接的に反映させるためにテキストをそれ自身の中に現れている名詞について分類することを考える。この意味での主題が照応現象とどう関わるか興味深い問題であるが本論文では触れない。

また加えて本稿では、主題抽出における文法情報の効果について実験結果を報告する。本研究では、特にハ、ガ、ヲ、ニなどの格助詞に注目した。コーパスの統計処理に文法情報を使うというのは (Hindle, 1990) の研究などに見られるが、テキストを語の集まりとして表すか、より複雑な言語学的な記述を使うか、という表現形式の問題は従来のテキスト分類の研究では懸案の一つになっている (Lewis, 1992)。

2 確率的主題推定

本節では主題を推定する方法ついて具体的に説明していく。まず、主題推定をテキスト分類として定義する。テキスト分類とは一般に、テキスト d が与えられたとき、予め用意された各分類項目 c について d に対する関連度 (条件つき確率) $P(c \mid d)$ を推定し、ある基準に基づいてテキストの分類項目を決定することである。主題推定はほぼ同様に定義されるが、ただ分類項目の扱いが異なる。主題推定では、分類項目としてテキスト内に出現した語を使う。したがって、主題推定とは、あるテキストに対してその中に出現している語で分類し、関連度の高い語をそのテキストの主題と呼ぼうというものである。なお、以下では分類項目を主題候補と呼ぶ。

テキスト d と主題候補 c の関連度 $\mathcal{L}(c \mid d)$ は以下の式で与えられるが、これは確率 $P(c \mid d)$ を適当な索引 t (index) について条件付けしたものにすぎない (Fuhr, 1989; Iwayama and Tokunaga, 1994)。(索引については第 3 節で触れる。) ただし、以下ではテキストを索引の集合と考える。索引はテキストに含まれる語でも、あるいは文法的な情報を付加したより複雑な構造であってよい。

$$\mathcal{L}(c \mid d) = \sum_t P(c \mid t) P(t \mid d)$$

さらに、ベイズの定理より、

$$\mathcal{L}(c \mid d) = P(c) \sum_t \frac{P(t \mid c) P(t \mid d)}{P(t)}$$

となる。$R(d)$ をテキスト d の索引集合とすると、

$$\mathcal{L}(c \mid d) = P(c) \sum_{w \in R(d)} \frac{P(T = w_i \mid c) P(T = w_i \mid d)}{P(T = w_i)} \quad (1)$$

となる。ここで、D を 訓練コーパスにおけるテキストの総数、D_c を D の中で c がタイトル (見出し) に現れたテキストの総数、F_c^* を D_c に現れた索引の総数、F_c^w を D_c 中の w_i の頻度、F_c^* を D_c の索引の総数、F_d^w をテキスト d における w_i の頻度、F_d^* をテストコーパス中のテキスト d の索引の総数、F_D^w を D における w_i の頻度、F_D^* を D の索引の総数とし、それぞれの確率を以下で推定する。

$$\begin{aligned} P(c) &= D_c/D \\ P(T = w_i \mid c) &= F_c^w / F_c^* \\ P(T = w_i \mid d) &= F_d^w / F_d^* \\ P(T = w_i) &= F_D^w / F_D^* \end{aligned}$$

すなわち、$P(c)$ は、主題候補 c が実際に主題である確率。$P(T = w_i \mid c)$ は、c が主題であるテキストに索引 w_i が現れる確率。$P(T = w_i \mid d)$ はテキスト d にお

ける w_i の出現確率である。また、$P(T = w_i)$ は無作為に選んだテキストに w_i が現れる確率である。

さて、(1) 式 では、特定の主題に関するテキストに特徴的（集中的）に現れている索引 w に重みが強くかかるようになっている。 これは基本的に情報検索 (Information Retrieval) でよく知られている $tf \cdot idf$ と呼ばれる語の重みづけ (term weighting) に相当する (Salton, 1988)。

3 索引表現

テキスト分類では、テキストは索引 (index) の集合として表現されるが、通常テキストを構成している語 (word) あるいは句 (phrase) が索引として使われる (Lewis, 1992; Finch, 1994)。

本研究では、索引として 2 つの表現形式を検討した。ひとつはテキストに現れている名詞で、もうひとつはその名詞に格情報を付加したものである。いづれの場合も、名詞とは形態素解析プログラム JUMAN が名詞と認定したものとした (Matsumoto et al., 1993)。また格情報の付加は同じ後置詞句内の名詞をそれぞれ格助詞でタグづけすることによって行った。ただし、後置詞句の解析は構文解析ではなく、修辞記号（句読点、終止符）を手がかりに大まかに自動で行った。格助詞の分類は (佐久間, 1983) に因った (表 1)。 実際のテストコーパスの解析例を図 1に示す。 名詞は '()' で、格助詞は '–'、格助詞のスコープは '<>' で示す。さらに、形態素解析の誤りを 1 つ星 ('⋆')、後置詞句解析の誤りを 2 つ星 ('⋆⋆') で示してある。 また、'ϕ' は空の格助詞で名詞が動詞の一部であることを示す。

テキストの索引はこのような解析済みコーパスから抽出する。 図 2に図 1から抽出した単純な単語ベースの索引表現と格助詞でタグづけした索引表現を示す。添字 'α' 'β' 'γ' 'δ' 'ϵ' はそれぞれ格助詞「ノ」、「ハ」、「ニ」、 「モ」に対応する。 ソシエテ$^{\beta}$、ジェネラル$^{\beta}$ は、実際は 'ソシエテ ハ' 'ジェネラル ハ' という索引表現に対応する。ちなみに、単語ベースとタグつきでは、若干違いがあることが分る。例えば、単語ベースでは 'キエフ' は 1 種類であるが、タグづきでは、'キエフ$^{\gamma}$' と 'キエフ$^{\alpha}$' というの 2 種類の索引表現ができる。

4 実験

前節で説明した 2 種類の索引法 (単語ベース vs. タグつき) を使い、主題推定の実験を行なった。以下ではその手続きと結果について報告する。

助詞	機能、意味
ガ	主格
ノ	所有、主格
ヲ	対格
ハ	提題
ニ	与格
ト	接続
デ	場所、手段、原因等
ヘ	方向
モ	提題
カ	並立
カラ	由来、出所
ヨリ	比較

表 1: 実験に用いた格助詞

4.1 手続き

実験では、92 年度ＣＤ－ＲＯＭ版日本経済新聞の 1 月から 6 月までの記事 42,401 件を選びだしコーパスとして利用した。そのうち同年 5 月 31 日までの記事 40,101 件を訓練用に 6 月 1 日以降の 2,000 件をテスト用に用いた。

グループ	記事サイズ (文字数)	記事数.
1	< 100	400
2	100-200	400
3	200-300	400
4	300-400	400
5	400-500	400

表 2: テストコーパス

さらに、テスト用データを長さ (文字数) 別に 5 つのグループに分けた (表 2)。 文字数 100 までの記事を第 1 グループ、100 以上 200 未満の記事を第 2 グループ、200 以上 300 未満の記事を第 3 グループ、300 以上 400 未満を第 4 グループ、400 以上 500 未満を第 5 グループに分類した。

訓練コーパスは、$P(c)$ $P(T = w)$ $P(T = w \mid c)$ を予め推定しておくのに用いた。$P(c)$ はテキストに主題 c が付与される確率であるが、実際にはテキストのタイトル (見出し) に現れた名詞をそのテキストの主題とした。したがって、$P(c)$ は名詞 c が記事の見出しに現れる確率である。 $P(T = w)$ は 見出し情報を除いた w の

(仏)(銀) が (キエフ) に (駐在)(員)(事務所)

＜(仏)(銀)(大手) の＞＜(ソシエテ)(ジェネラル) は＞＜(15 日 **)、(ウ *)(クラ *)(イナ *) の＞＜(首都)(キエフ) に＞＜(駐在)(員)(事務所) を＞　＜(開設) ϕ ＞すると＜(発表) ϕ ＞した。すでに＜(キエフ)(市) (当局) の＞＜(許可) を＞得たと言う。

図 1: 記事の解析例

$R(d) =$ { 仏, 銀, 大手, ソシエテ, ジェネラル, 15 日, ウ -, クラ -, イナ, 首都, キエフ, 駐在, 員, 事務所, 開設, 発表, 市, 当局, 許可 }

$R'(d) =$ { 仏$^{\alpha}$, 銀$^{\alpha}$, 大手$^{\alpha}$, ソシエテ$^{\beta}$, ジェネラル$^{\beta}$, 15 日$^{\alpha}$, ウ -$^{\alpha}$, クラ -$^{\alpha}$, イナ$^{\alpha}$, 首都$^{\gamma}$, キエフ$^{\gamma}$, 駐在$^{\delta}$, 員$^{\delta}$, 事務所$^{\delta}$, 開設$^{\phi}$, 発表$^{\phi}$, キエフ$^{\alpha}$, 市$^{\alpha}$, 当局$^{\alpha}$, 許可$^{\epsilon}$ }

図 2: 索引言語： 単語ベース $R(d)$ vs. タグつき $R'(d)$

出現確率、$P(T = w \mid c)$ は見出し語 c の下での見出し情報を除いた w の出現確率である。

4.2 結果

以下では主題推定モデル (式 1) のパフォーマンスについて報告する。評価は前節の 5 つのテストグループについてそれぞれ索引言語として $R(d)$ (単語ベース) を用いた場合と $R'(d)$ (タグつき) を用いた場合で行なった。

主題推定は 2 段階でおこなう。まず、式 1 により 主題候補 c とテキスト d の関連 (連想) 度 $\mathcal{L}(c \mid d)$ を求め、次に適当な決定戦略 (decision rule) に基づきテキストに主題を付与する。今回の実験ではとりあえずテキストに現れている名詞をすべて主題候補と考えることにした。つまり、$c \in R(d)$ である。

テキスト分類では決定戦略として、以下が代表的である。

- **k-per doc :** 同じテキストについて、主題候補 $c_1 \ldots c_N$ の中から連想度の高い順に上から k 個とる。
- **probabilistic thresholding :** ある閾値 s を決めて、$\mathcal{L}(c \mid d) \geq s$ なる c をすべてとる。
- **proportional assignment :** 主題候補 c について、$\mathcal{L}(c \mid d_1) \ldots \mathcal{L}(c \mid d_N)$ の上位から訓練データでの c の付与率に合せてとる。(例えば、訓練データで c の付与率が全体で 2% であれば、テストデータ N の上位 4%、8% あるいは 10% をとる。)

今回の実験では、2 番目の probabilistic thresholding を決定戦略として採用した。k-per doc 方式は一般に他の方式に比べて分類の精度が悪いことが知られている。また、今回の実験では時事性の強い新聞を利用したため、訓練コーパスでの主題候補の付与率がテストコーパスの主題付与に反映するとは考えにくい。そのため、proportional assignment 方式は考えなかった。

(Gale et al., 1992) に従い、ベースラインとの比較を行なった。ベースラインは課題の難易度を計る 1 つの考え方であり、精度の下限を決めるのに用いられる。特に定義はないが、(Gale et al., 1992) によれば、

> "The baseline represents a simple, straw man approach to the task, which should be outperformed by any reasonable model."

ということになる。今回の実験ではテキストから無作為に選んだ名詞が見出しに現れる確率 (つまり、$P(c \mid d)$) をベースラインとした。

グループ	文書サイズ (文字数)	タグ無	タグ付	ベースライン	補正	平均方式
1	< 100	49%	54%	19%	unscaled	micro
2	100 - 200	42%	44%	33%	unscaled	micro
3	200 - 300	35%	37%	30%	unscaled	micro
4	300 - 400	31%	32%	32%	unscaled	micro
5	400 - 500	31%	33%	35%	unscaled	micro

表 3: 結果のまとめ

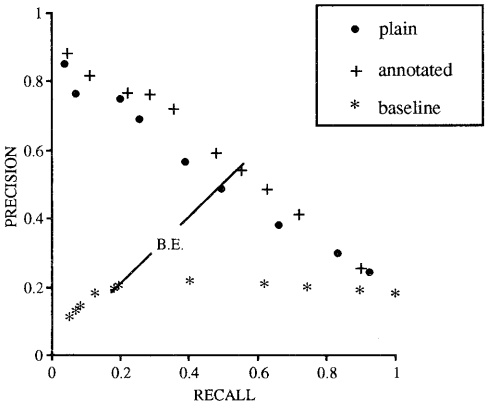

図 3: 主題推定の精度 (テキストサイズ 100 文字未満)

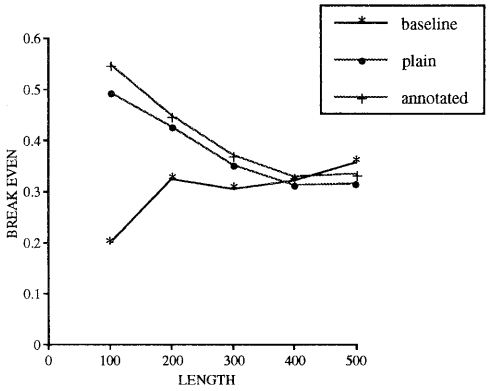

図 4: テキストサイズ (横軸) に対する精度の変化

さて、結果であるが、まず 100 文字未満 (2 ～ 3 文程度) のテキストについて見てみることにする (図 3)。図で 'plain' は索引言語として $R(d)$、'annotated' は $R'(d)$ を選択した場合である。また、'B.E.' は、break-even point(引き分け点、recall=precision で最も高い点) を表す。

図 3では、B.E. で $R'(d)$(54%) が $R(d)$(49%) を 5% 程度上回っており、格情報の効果は明らかである。いづれの場合も、ベースライン (19%) を凌いでいる。

次にテキストのサイズに対するモデルの精度の変化を調べてみた (図 4、表 3)。なお、recall/precision は micro-average 方式 (Lewis, 1992) で平均し、$\mathcal{L}(c \mid d)$ に対する正規化は行わなかった (unscaled)。図 4が示すように、テストを行った 5 つのグループいづれでも $R'(d)$ が勝っているが、テキストのサイズが大きくなるにつれ次第にモデルの精度が劣化し、文字数 400 ～ 500 あたりでベースラインと逆転しているのが分る。原因はテキストのサイズが増大するにつれ索引言語、主題候補共にノイズが多く含まれるようになったためであると考えられる。解決するには何らかの方法で索引言語、主題候補を絞りこむことが必要になってくる。相互情報量 (mutual information) (Hindle, 1990) のアイディアを使うのも 1 つの方法であるが、いまのところ有効性については未確認である。

参考までに、図 1に現れた名詞の主題性を式 1 に基づいて推定してみた (表 4)。'ソシエテ' 'ジェネラル' 'キエフ' は訓練セットの見出しに出現しなかったので、推定は不可能であった。上位に記事の見出し語が並んでいるのがわかる。

5 まとめ

本稿では、テキストの主題を自動的に抽出するための手法を提案した。本手法の特徴は主題推定をテキスト分類と捉え、テキストを自身の中に現れている名詞で分類したところにある。本方式の有効性を検証するため実際のコーパス (92 年度日本経済新聞ＣＤ－ＲＯＭ版、記事 42,401 件) を使って実験を行った。実験のポイントは、(1) 主題推定のモデル (式 1) の性能と (2) 文法情報が主題推定に影響を及ぼすかどうかという点である。

名詞	$\mathcal{L}(c \mid d)$
クラ -	0.075
ウ -	0.074
イナ	0.073
仏	0.057
銀	0.045
事務所	0.032
開設	0.018
市	0.017
首都	0.011
員	0.008
発表	0.005
駐在	0.004
許可	0.0029
大手	0.0027
当局	0.0026
15 日	0.0001
ソシエテ	—
ジェネラル	—
キエフ	—

表 4: 各単語における話題性の度合

結果は、(1) に関して、長さ 300 文字未満の記事ではベースラインを上回ることが確認された (表 3)。しかし、300 文字以上についてはベースラインと同等あるいは下回ることが分かった。 原因としては主題を $R(d)$ から選択しているため長いテキストでは当然のことながら候補数が多く、そのため精度が劣化したと考えられる。したがって、精度を維持するには主題候補の数を抑える必要がある。(2) については、ポジティブな結果が得られた。テストした 5 つのコーパスセットにおいて 1% ～ 5% の幅で格情報の導入が精度向上につながった。

参考文献

Finch, S. (1994). Exploiting Sophisticated Representations for Document Retrieval. In *Proceedings of the Fourth Conference on Applied Natural Language Processing.*

Fuhr, N. (1989). Models for Retrieval with Probabilistic Indexing. *Information Processing & Management*, 25(1):55–72.

Gale, W., Church, K. W., and Yarowsky, D. (1992). Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs. In *Proceedings of American Association for Computational Linguistics*, pages 249–256.

Grosz, B. and Sidner, C. (1986). Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.

Hindle, D. (1990). Noun Classification from Predicate-Argument Structures. In *Proceedings of the 22nd Annual Meeting of the Association of Computational Linguistics.*

Iwayama, M. and Tokunaga, T. (1994). A Probabilistic Model for Text Categorization: Towards a Tool for Personal Knowledge Aquisition. In *Proceedings of the Fourth Conference on Applied Natural Language Processing.*

Kuno, S. (1973). *The Structure of the Japanese Language.* The MIT Press, Cambridge, Mass.

Lewis, D. D. (1992). An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–50.

Matsumoto, Y., Kurohashi, S., Utsuro, T., Taeki, Y., and Nagao, M. (1993). *Japanese Morphological Analysis System JUMAN Manual.* Kyoto University. In Japanese.

Salton, G. (1988). *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, Reading, MA.

Sidner, C. L. (1983). Focusing in the comprehension of definite anaphora. In Brady and Berwick, editors, *Computational Model of Discourse*, pages 267–330. The MIT Press, Cambridge.

Walker, M., Iida, M., and Cote, S. (1994). Japanese Discourse and the Process of Centering. *Computational Linguistics*, 20(2):193–232.

佐久間, 鼎. (1983). 現代日本語の研究. くろしお出版.