

潜在意味解析を用いた新聞記事の見出し生成

Estimating Headlines Using Latent Semantics

奥村 直也[†] 三浦 孝夫[†]

[†] 法政大学理工学研究科 〒184-8584 東京都小金井市梶野町 3-7-2

E-mail: [†]naoya.okumura.7s@stu.hosei.ac.jp, ^{††}miurat@hosei.ac.jp

あらまし 本研究では、新聞記事の見出し生成手法を提案する。新聞記事の見出しは、特徴的な表現（用語や言い回し）で記述されていることが多い。従来手法では、記事本文からのキーワード抽出やパターン抽出手法が提案されている。しかし、特徴的な表現を考慮することができないという問題がある。本研究では記事内容の空間と見出し表現の空間を分け、類似した記事は類似した見出しが対応すると仮定して、その対応を考察する。潜在意味解析を利用することにより、類似した記事を選択し、その記事の見出しの類似性を与える。

キーワード 潜在意味解析, 見出し生成

1. ま え が き

近年、インターネットで電子書籍や新聞記事が容易に入手できるが、何ら処理されることなく流れ去っていることが多い。これらの情報には内容を表した見出しが付与されている。見出しは、記事本文の特徴を表現したものである。そのため、記事本文の要約とは違い、記事の見出しには、特徴的な表現（または単語）で記述されていることが多い。

しかし、こういった見出し表現を抽出することは困難である。情報検索や機械学習分野での従来手法では、記事からの発生頻度・共起性・エントロピに基づくキーワード抽出や、決められた品詞・名詞の並びのパターンに基づくパターン抽出手法が提案されている。これらの手法では、要約にはなるが、特徴的な表現を得ることが難しく、一方で内容の類似性を判定する必要がある。

本研究では、類似した記事内容は類似した見出し表現がなされていると仮定する。記事内容の類似性を推定するために、潜在意味解析 Latent Semantic Analysis (LSA) を利用する。潜在意味解析とは、文書集合とその文書に含まれる単語の共起性にもとづいて潜在意味を抽出し、各単語を潜在意味のベクトルで表現し、その関係を分析する技術である。本研究では、潜在意味解析を利用して、類似した新聞記事を選択し、その類似した記事の見出しを使用することで新たな見出しを生成する。2章では記事内容の表現と見出しの表現の概要を述べ、3章では本文検索についてを述べる。4章では特徴語抽出についてを述べる。5章では実験を行い、得られた結果に関する考察を述べ、6章を結論とする。

2. 記事内容の表現と見出しの表現

2.1 記事の表現

記事本文を検索や解析するためには、単語による表現方法である BagOfWords モデルと潜在意味解析が用いられている。BagOfWords モデルは、単語を特徴量の集合とみなし、出現

頻度 TF (Term Frequency) や TF・IDF (Inverse Document Frequency) などの重みを特徴量として表現するモデルである。潜在意味解析は、文書集合とその文書に含まれる単語を潜在意味の集合で表現する方法である。

BagOfWords モデルは文書全体を表現し、潜在意味解析は文書の主張を表現している点で異なる。さらに、BagOfWords モデルでは、類義語や同義語などの類似した語を区別することができない。

潜在意味解析は、単語の概念として潜在意味を用いるため、潜在意味が類似していれば、類義語や同義語を区別することができる。これらの BagOfWords モデルや潜在意味表現などの個々の特徴量を1つのまとまった表現をするために、ベクトル空間モデルが用いられる。

2.2 見出しの表現

記事の見出しは、本文とは違い、明確な構文形式を持たず、特徴的な表現（単語や言い回し）で記述されている。見出しを生成するために、キーワードや要約として文章を抽出する手法があるが、記事内の文章を引用することは少ない。例えば、（ファイル：きづなの質問「与党枠で」）という見出しで、記事本文が以下のようになっている。

本文：（自民党の岸田文雄国対委員長は30日、民主党に離党届を出した9議員でつくる新党きづなについて、「野党枠で（質問する）というのは遠慮したい」として、衆院予算委員会での質問時間の配分にあたっては与党の持ち時間を割り振るよう、民主党の城島光力国対委員長に申し入れた。）

キーワードとして単語（きづな、与党、質問 など）は抽出できるが、見出し表現には類似しない。そのため、記事からではなく、記事の見出しから見出し表現を抽出しなければならない。

2.3 潜在意味抽出

潜在意味解析の基本となる主成分分析 (PCA) は、データとなる文書の見方を変えて主成分と呼ばれる特徴的部分（意味単位）で表現したものである。主成分分析では、主成分で物事を

考えると特徴が浮き彫りになっているため、理解がしやすくなる点が重要である。この利点を受け継ぐものが潜在意味解析である。潜在意味解析では、主成分を潜在意味と呼び、2 項の軸（単語、記事）に対してそれぞれの潜在意味（主成分）で表現することができる。これらの潜在意味は、ベクトル表現により、1 つの特徴量として表現することができる。

3. 本文検索

記事は本文と見出しからなる。記事の見出しを推定するには、他の見出しを利用する。類似した記事は、類似した見出しを持つと考えられるため、類似記事の検索を行う。記事類似性を判定するために、潜在意味解析を用いる。見出しなし未知記事に対して推定を行う。

本文検索は以下のように示す。未知記事 q : 1 件と見出しあり既知記事: k 件 ($k > 1$) を準備する。潜在意味解析を行うために、既知記事集合 D に対して特異値分解 (SVD) を行う。特異値分解は、潜在意味で表現された単語・単語行列 U 、単語と文書の潜在意味行列 S 、潜在意味で表現された文書・文書 V に分解できる。単語・文書行列 D の特異値分解は、以下の式で表される。

$$D = U \times S \times V^t$$

次に質問記事 q を検索するために、潜在意味空間へ射影する。潜在意味空間へ射影すると以下の式になる。

$$q^t U (S^{-1})^t$$

射影した q と既知記事集合のそれぞれの文書を潜在意味で表現された V を検索する。行列内のすべての記事と類似度を計算する。射影した q と V の中の文書 i の類似度は以下のように定義される。

$$\cos \theta_i = \frac{(q^t U (S^{-1})^t, V_i)}{|q^t U (S^{-1})^t| |V_i|}$$

類似度の高い順に q の候補記事として抽出できる。候補記事の見出しが q の見出し候補となる。

4. 特徴語抽出

見出しは記事の内容を特徴付ける表現であるから、その自立語はすべて（記事の）特徴語である。見出しは特徴語の配置を扱う枠組であるとみなせば、本文には特徴語を述べる表現を含むであろう。逆に、そのような表現は記事を表す新たな特徴語を定義している。この考え方に基づけば、特徴語 W は見出しに出現するか、本文内で定義され、本文内で定義される語は記事の特徴付ける可能性が高い。特徴語 W は、従って (1) 見出しに生じる自立語、(2) 本文に定義されている自立語、のいずれかであると仮定する：

$$char(W) \leftarrow head(W)$$

$$char(W) \leftarrow body(EX(W))$$

式 $body(EX(W))$ は W が n -gram $EX(Y)$ の位置 Y に出現する語であることを示す。逆に特徴語 W がこのような出現パターン $EX(Y)$, $Y = W$ を有すれば、この式により新たな特徴語を産む可能性がある。本稿では、パターンとは、「単語の並

び」や「名詞のみの並び」などのことを表している。これにより、キーワードや特徴的な並びのフレーズを抽出することができる。見出しに出現する（代名詞や軽動詞を除いた）自立語はすべて特徴語とする。特徴語 X の前 4 個、後 4 個以下の自立語列をパターン $EX(X)$ とし、 $EX(W)$ の間に含まれるすべての語 W を特徴語とする。抽出された中から、名詞のみを特徴語として扱う。新しく特徴語に加えられた単語を含め、新たにパターンを探し、特徴語をさらに抽出する。以降この繰り返し、新たな特徴語が抽出できなくなった時点で終了する。

具体的な例として、表 1 の記事について考える。head(W) を特徴語として抽出する。表 1 の head(W) の単語が抽出でき、char(W) に加える。次に char(W) からパターンを探す。表 1 のパターンが存在する。このパターンに当てはまる単語を抽出する。この単語（ダルビッシュ）以外にパターンに当てはまる単語は、body($EX(W)$) の単語（投手）である。これを char(W) に加える。以降、char(W) から新たなパターンを抽出し、body($EX(W)$) を推定する。body($EX(W)$) が新たな単語を得られなければ終了となる。表 1 から最終的に得られた特徴語 char(W) は、単語（絶対、エース、ダルビッシュ、チーム、チャンス、シーズン、判断、日本ハム、投手）である。

表 1 特徴語抽出の例題

本文
絶対的なエースとして君臨してきたダルビッシュがチームを離れる見込みだが、「のチームにもチャンスが生まれてくるシーズン。どうやって生かすか、一つの判が大事になる。そこだけは間違わないようにしたい」と意気込んだ。エースとして臨してきた投手がチームを離れるのは、日本ハムとしても大きな痛手になる。
head(W)
絶対、エース、ダルビッシュ、チーム、チャンス、シーズン、判断、日本ハム
パターン
エース、君臨する、(ダルビッシュ)、チーム、離れる
body($EX(W)$)
投手
char(W)
絶対、エース、ダルビッシュ、チーム、チャンス、シーズン、判断、日本ハム、投手

5. 実験

本稿は実験環境と評価方法を述べ、実験により得られた結果を示す。得られた結果に関する考察を述べる。本研究では、本文検索と特徴語抽出の二つの実験を行う。本文検索では、選択した見出しが正解見出しと類似しているものを選択できたか検討する。特徴語抽出では、記事に合った特徴語を抽出できたかと特徴語として正しいかを検討する。

5.1 実験環境と評価

実験に使用するコーパスは、毎日新聞記事データ集 2012 より 1 月から 6 月分の半年分の記事 17615 件を用いる。質問記事として、全体の新聞記事 17615 件から記事 1761 件を使用する。形態素解析ソフト Mecab を用いて、形態素解析を行い、文書毎に名詞を抽出する。学習に使用した記事の見出しから抽出できた名詞は 12197 語である。ベースラインは、特徴語抽出では、質問記事の名詞を出現頻度 TF で表現し、ランキングを取り、特徴語を抽出を行う。本文検索では、すべての記事を TF で表現し、余弦類似度を取り、本文を検索する。

本文検索と特徴語抽出の実験についてそれぞれ行う。本文検索では、選択した見出しが正解見出しと類似しているものを選

択できたか検討する。正解見出しは、質問記事に対して、本来のコーパスなしで存在している見出しをこの質問の「正解」という。(1) 質問記事に正解見出しを用意し、質問記事と類似する記事を選択する。提案手法では、質問記事を潜在意味ベクトルに変え、潜在意味解析を用いて記事を選択する。選択する記事を潜在意味ベクトルによる余弦類似度で上位 10, 20, 30, 40, 50 件まで抽出する。その選択した記事の見出し (50 件) と正解見出しに生じる自立語が、余弦類似度で 30%以上のものが 1 件でも上位 10, 20, 30, 40, 50 件までに含まれていれば正解とする。質問記事それぞれに対して同様に行い、質問記事全体がどれだけ上位 10, 20, 30, 40, 50 件で正解したかを適合率として調べる。

特徴語抽出では、見出しに合った特徴語を抽出できたかと特徴語として正しいかを検討する。(2) 質問記事に対して、本文中に含まれる特徴語を抽出する。このとき、これは見出しに生じていれば正しい特徴語とし、本文中に生じる特徴語で見出しに生じている割合を適合率という。(3) 質問記事の見出しに生じる特徴語すべてが、その本文中に特徴語として生じている割合を再現率という。本論文では、特徴語の適合率・再現率を評価対象とする。

5.2 実験結果

見出し推定で得られた結果を表 2 に示す。

表 2 見出し推定の適合率 (1)

ランキング	提案手法	TF
10	29.04	21.43
20	56.79	41.28
30	79.09	52.16
40	90.32	67.33
50	92.31	72.12

特徴語抽出で得られた結果を表 3 に示す。

表 3 特徴語抽出の適合率 (2) と再現率 (3)

適合率	提案手法	TF
	57.37	40.19
再現率	提案手法	TF
	62.41	52.81

本文検索の場合、適合率を求めると表 2 のようになり、提案手法と TF を比べると最大 26.93%の向上が見られる。特徴語抽出の場合、まず、抽出できた特徴語 10 と抽出できたパターン 11 が示してある。抽出できた特徴語では、頻度 2 と頻度 3 のみで 96.7%，抽出できたパターンでは、頻度 2 と頻度 3 のみで 99.7%を占めている。さらに、適合率を求めると表 3 のようになり、提案手法と TF を比べると 17.18%の向上が見られ、再現率は 9.6%の向上が見られる。

5.3 考察

表 2 より、本文検索では、適合率を比較すると、ランキング 10, 20, 30, 40, 50 件まで選択した見出しの適合率はベースラインよりも提案手法がすべて優る結果となり、的確に本文検索を行えていると考えられる。提案手法が優っている場合、表 4 の結果から、ランキング 15 件までの提案手法が 12 件、ベースラインが 1 件と正解数も多く、最大適合率 80%となる。これは、潜在意味で検索行う方が優れていることを示している。さ

らに、ランキング 30 件までで 26.93%(=79.09-52.16) の向上が見られる。また提案手法が劣っている場合、表 5 の結果から、ランキング 20 件までの提案手法が 0 件、ベースラインが 0 件と正解する見出しが存在しない。主な理由は、この質問記事が社説の記事なので、本文に特徴語がほとんど現れず、他の社説記事も同様に特徴的でないためと考えられる。

特徴語抽出は、再現率・適合率ともに提案手法が優れている。表 6 をみると、正解見出しに生じる 8 単語 (ヤング, フェスティバル, 千葉, 絆, 地域, 人, たち, 交流) 中の抽出できた特徴語 17 単語が選ばれ、適合率は 35.3%(6/17), 再現率は 75.0%(6/8) となる。特徴語は、固有名詞として抽出した単語 (千葉, ヤング), 他の見出し語を利用される抽出される単語 (県 交流, 青少年, グルメ, コーナー, 町, 人, ヤング, フェスティバル, 昨年, 企画, 高, たち, 入場), パターンを利用して抽出される単語 (協会) で構成される。正解として大きく影響与えているものが、他の見出し語を利用して抽出される単語である。元々見出し語として他の見出しに選ばれている分、類似した内容の記事が記事集合にあると大きく影響が出る。記事固有の単語を抽出するために固有名詞を抽出した結果、頻度の高い固有名詞は、見出し語に含まれやすいのではないかと考えられる。固有名詞として抽出される単語 (千葉, ヤング) は頻度 2 回以上出現している単語である。他の例では、表 7 をみると、正解見出しに生じる 15 単語, 見出し抽出できた特徴語が 27 単語選ばれ、適合率は 25.9%(7/27), 再現率は 46.7%(7/15) となる。

特徴語の劣っている適合率と再現率の例は以下となる。表 8 では見出しの特徴語が 6 単語 (視聴, 室, 湯けむり, スナイパー, お正月, SP), 抽出できた特徴語は 3 単語 (屋, さん, ワケ) である。適合率と再現率ともに 0%となる。他の例では、表 9 では見出しの特徴語が 10 単語 (シンポジウム, 作家, 越境, テーマ, 日本, 大, リービ, 英雄, さん, 基調, 講演), 抽出できた特徴語は 6 単語 (者, 日本語, 教授, 作家, 賞, 受賞, シンポジウム) である。適合率 28.6%と再現率 20.0%となる。

6. 結論

本稿では、潜在意味解析を用い、類似した本文を選択し、その記事の見出しを利用することで新たな見出し推定を提案した。実験では、提案手法として本文検索と特徴語抽出を行ったが、両方共にベースラインに優る結果となり、見出しと特徴語を組み合わせることにより、適合する見出しを提示できた。また、提案手法で本文検索の結果が最も 30 件までの検索の場合、向上率 17.18%と高く、潜在意味解析により選択された見出しが正解の見出しと類似していた。特徴語抽出の結果が、適合率が 16.83%, 再現率が 9.6%の向上が見られた。

文 献

[1] Das, D. and Martins, A.F.T: A Survey on Automatic Text Summarization Tech. Report, Univ. of Duisburg-Essen, 2007
[2] Kleinbaum, D.G. and Klein, M.: Logistic Regression: A Self-Learning Text (Statistics for Biology and Health),

表 4 見出し推定の正解例

正解見出し：ブックウォッチング：インフォメーション 早読みシートで読書を楽しむ		
ランキング	提案手法から選択された見出し	正解
1	ブックウォッチング：インフォメーション 「子どもの暮らし 生きる知恵」 斉藤洋講演会	○
2	ブックウォッチング：インフォメーション 矢田亜希子さん、よみかかせ会を開催	○
3	ブックウォッチング：新刊 『夜の国のクーバー』＝伊坂幸太郎・著	○
4	ブックウォッチング：インフォメーション おいしく楽しい食生活応援の健康レシピ集	○
5	ブックウォッチング：新刊 『テキスタイル用語辞典』＝成田典子・著	○
6	今週の本棚：沼野充義・評 『おろか者たち—中学生までに読んでおきたい哲学（4）』＝松田哲夫・編	×
7	ブックウォッチング：街の本屋さん 旅の本屋 のまど（東京都杉並区）	○
8	今週の本棚・情報：「名歌名句大事典」を刊行	×
9	ブックウォッチング：新刊 『いかにして問題をとくか 実践活用編』＝芳沢光雄・著	○
10	今週の本棚：中村達也・評 『世代間格差—人口減少社会を問いなおす』＝加藤久和・著	×
11	ブックウォッチング：新刊 『愛されるアイデアのつくり方』＝鹿毛康司・著	○
12	ブックウォッチング：新刊 『大転換 「BOP」ビジネスの新潮流』＝田原総一郎・著	○
13	ブックウォッチング：インフォメーション 「西尾幹二全集」刊行記念講演会を開催	○
14	ブックウォッチング：新刊 『「三畳小屋」の伝言 陸軍大将 今村均の戦後』＝朝野富三・著	○
15	ブックウォッチング：新刊 『大澤真幸 THINK「O」第10号記念号』＝大澤真幸、岩井克人・著	○
16	今週の本棚：持田叙子・評 『南方熊楠大事典』＝松居竜五、田村義也・編	×
17	今週の本棚・本と人：『負けんとき ヴォーリス満喜子の種まく日々 上・下』 著者・玉岡かおるさん	×
18	今週の本棚：沼野充義・評 『人生と運命 全3巻』＝ワシーリー・グロスマン著	×
19	今週の本棚・新刊：『からのゆりかご 大英帝国の迷い子たち』＝マーガレット・ハンフリーズ著	×
20	今週の本棚・新刊：『京都洋館ウォッチング』＝井上章一・著	×
21	ブックウォッチング：インフォメーション 総合エンターテインメント文芸誌「読楽」創刊	○
22	今週の本棚・情報：岩波文庫を電子配信	×
23	今週の本棚・新刊：『遣隋使がみた風景 東アジアからの新視点』＝氣賀澤保規・編	×
24	ブックウォッチング：新刊 『おほソとの対話』＝高田好胤・著	○
25	今週の本棚・新刊：『イタリア文化事典』＝イタリア文化事典編集委員会・編	×
26	ブックウォッチング：新刊 『西本願寺への誘い 信仰がまもり伝えた世界文化遺産』＝岡村喜史・著	○
27	今週の本棚：山崎正和・評 『安部公房の都市』＝菊部直・著	×
28	今週の本棚：松原隆一郎・評 『食の終焉』＝ポール・ロバーツ著	×
29	今週の本棚：本村凌二・評 『前キリスト教的直観—甦るギリシア』＝シモヌ・ヴェイユ著	×
30	ブックウォッチング：新刊 『降霊会の夜』＝浅田次郎・著	○
	提案手法により抽出された特徴語	
	こと、スピード、本、時間、発売、学習	
ランキング	ベースラインから選択された見出し	正解
1	今週の本棚：若島正・評 『別名S・S・ヴァン・ダイン—ファイロ・ヴァンスを創造した男』＝ジョン・ラフリー著	×
2	今週の本棚：張競・評 『「文化力」の時代—21世紀のアジアと日本』＝青木保・著	×
3	今週の本棚：白石隆・評 『中国共産党—支配者たちの秘密の世界』＝リチャード・マクレガー著	×
4	今週の本棚：富山太佳夫・評 『はるかなる岸辺』／『ニグロとして生きる』	×
5	今週の本棚・新刊：『平成不況の本質 雇用と金融から考える』＝大瀧雅之・著	×
6	今週の本棚・新刊：『ベルマの独裁者 タンシュエ』＝ベネディクト・ロジャーズ著	×
7	今週の本棚・新刊：『放射線医が語る 被ばくと発がんの真実』＝中川恵一・著	×
8	今週の本棚・情報：小学高学年からの現代史	×
9	今週の本棚・情報：『日本政党史論』を復刊	×
10	今週の本棚・本と人：『こども東北学』 著者・山内明美さん	×
11	今週の本棚・新刊：『世界記録はどこまで伸びるのか』＝ジョン・ブレンカス著	×
12	今週の本棚・新刊：『震災トラウマと復興ストレス』＝宮地尚子・著	×
13	ブックウォッチング：インフォメーション ジョブズ氏の生涯を漫画で日本初刊行	○
14	今週の本棚・MAGAZINE：『住む。sumu』2012年冬号	×
15	今週の本棚・MAGAZINE：『真夜中 No. 15』	×
16	今週の本棚・新刊：『眼の海』＝辺見庸・著	×
17	今週の本棚・新刊：『山田方谷の夢《文庫版》』＝野島透・著	×
18	今週の本棚・今週の執筆者：荒俣宏さんほか	×
19	ブックウォッチング：インフォメーション 「西尾幹二全集」刊行記念講演会を開催	○
20	ブックウォッチング：インフォメーション 「家庭菜園検定」の参考テキストを発売	○
21	ブックウォッチング：新刊 『弁護士探偵物語 天使の分け前』＝法坂一広・著	○
22	今週の本棚・今週の執筆者：小林信彦さんほか	×
23	今週の本棚：中村桂子・評 『親切的進化生物学者』＝オレン・ハーマン著	×
24	今週の本棚：三浦雅士・評 『世界文明史の試み—神話と舞踊』＝山崎正和・著	×
25	今週の本棚：池内紀・評 『旅と観光の年表』＝旅の文化研究所・編	×
26	ブックウォッチング：新刊 『この国の環境 時空を超えて』＝陽捷行・文、ブルース・オズボーン写真	○
27	ブックウォッチング：新刊 『もしもあなたががんになったら』＝大岩孝司・著今週の本棚：辻原登・評 『野いばら』＝梶村啓二・著	○
28	今週の本棚・新刊：『梅棹忠夫の「人類の未来」』＝梅棹忠夫・著、小長谷有紀・編	×
29	今週の本棚・新刊：『世界軍歌全集』＝辻田真佐憲・著	×
30	今週の本棚：第10回毎日書評賞の贈呈式 海部宣男さん、喜びの声代読で	×
	ベースラインにより抽出された特徴語	
	BRLM、スピード、本、こと、時間、好き	

Springer-Verlag, 2010

[3] 奥村 学, 難波 英嗣: テキスト自動要約 (知の科学)

[4] NICT Information Analysis Laboratory, National Institute of Information and Communications Technology, Japanese WordNet, <http://nlpwww.nict.go.jp/wn-ja/index.en.html>

[5] 大澤 幸生, ベネソン ネルス E, 谷内田 正彦: KeyGraph：語の共起グラフの分割・統合によるキーワード抽出

[6] Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004, May). WordNet:: Similarity: measuring the relatedness of concepts. In Demonstration Papers at HLT-NAACL 2004 (pp. 38-41). Association for Computational Linguistics

[7] Popescu, A. and Ungar, L.H.: Automatic Labeling of Document Clusters, KDD-2000

[8] Saranyamol, C.S. and Sindhu, L.: A Survey on Automatic Text Summarization International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 5 (6), 2014, pp. 7889-7893

[9] Yuen-Hsien Tseng, Chi-Jen Lin, Hsiu-Han Chen, Yu-I Lin: Toward Generic Title Generation for Clustered Documents, Third Asia Information Retrieval Symposium, AIRS 2006, Singapore, October 16-18, 2006, pp.145-157

表 5 見出し推定の不正解例

	正解見出し； 正解見出し：社説：２０１２・激動の年 復興へ本格的な歩みを	
ランキング	提案手法から選択された見出し	正解
1	ファイル：対中外交政策を提言	×
2	ファイル：東日本大震災 がれき処理、国有林活用も検討	×
3	首相日々：３０日	×
4	原子力規制庁：きょう設置法案審議入り 自民も一転出席へ	×
5	ファイル：「会見で丁寧に」閣議決定	×
6	ファイル：改正沖縄２法、月内成立へ	×
7	松田喬和の首相番日誌：首相の政治生命力	×
8	衆院選挙制度改革：各党反発、混迷深まる 民主、解散先送りに好都合	×
9	社告：毎日フォトバンク 【大阪】	×
10	ファイル：新党きづなが衆院新会派	×
11	輿石・民主幹事長：一体改革か、元代表か 見えぬ真意	×
12	ファイル：行革推進、有識者懇新設へ	×
13	ファイル：「新党大地」が参院新会派	×
14	原子力規制庁：与党が譲歩 自公案のみ、委員会「独立」 ７月にも設置	×
15	首相日々：３１日	×
16	首相日々：８日	×
17	社告：「書の甲子園」 力作２１３４点展示―大阪であすまで 【大阪】	×
18	ファイル：女性宮家「１代限りが最大公約数」―第４回ヒアリング	×
19	社説：夏の節電対策 脱原発社会への一歩に	×
20	消費増税法案：きょう野田・小沢会談 首相、説得か決別の岐路 元代表、譲歩せず長期戦狙う	×
21	原子炉等規制法改正案：原発４０年廃炉 例外規定で形骸化も 安全確保、専門家「公的検査を」	×
22	社説：平田容疑者逮捕 逃亡の実態解明が必要	×
23	ファイル：前原・民主政調会長、産経記者の出席認める	×
24	ファイル：待機児童、４年ぶり減少	×
25	社告：読むひろの世界 １５、２１日子ども向け「おはなし会」―兵庫県立美術館 【大阪】	×
26	社告：新進気鋭の音色を 日本音コン受賞者演奏会、堺で４月７日 【大阪】	×
27	松田喬和の首相番日誌：「９３年型」か「０５年型」か	×
28	ファイル：モンゴル軍訓練施設視察―川防衛相	×
29	北朝鮮：ミサイル発射予告 田中防衛相、米と協調確認へ	×
30	社説：東京スカイツリー 都市を考える新名所に	×
	提案手法により抽出された特徴語	
	温暖、使用、大量、的、者、送電、高台、サイクル、県、中間、停止、大手、電力、対策、事故、年間、化、分離、自治体、稼働、全国、住宅、供給、計画、被災、地域、性、住民、中、地、原発、価格、網、経済、汚染、負担、よう、再生、燃料、核、制度、安定、費、促進、福島、移転、道筋、発送、先月、立地、課題、事業、最終、量、発電、必要、依存、染、新規、料金、避難、問題、以外、難題、責任、帰還、エネルギー、可能、除、参入、処理、処分、確保、電気、区域、特区、声、設置、解除、生活、こと、復興、年、安全、仮設、貯蔵、被害、基本、政策、	
ランキング	ベースラインから選択された見出し	正解
1	みんなの広場：役に立たなかった２０００円札＝無職・十河隆夫・７０	×
2	税と社会保障：一体改革 民主、法案反対で除籍例なし	×
3	ファイル：自民、民主とは政策連携で	×
4	主な選挙：６月の主な選挙結果	×
5	社説：認知症と福祉 地域の受け皿なければ	×
6	みんなの広場：今夏は原発不要を実証しよう＝無職・濱崎和子・７７	×
7	社説：視点・「仮の町」構想 ニュータウン方式は疑問＝論説委員・人羅格	×
8	社説：年寄名跡問題 未来の角界担える人を	×
9	社説：リオプラス２０ 緑の経済へと進めよう	×
10	みんなの広場：カカオ生産地の子供に思いを＝大学生・末次志穂・１９	×
11	みんなの広場：家での散髪も味があるが…＝理容師・服部直記・６４	×
12	みんなの広場：水俣病と原発事故は似た構図＝無職・鷺柳一・６５	×
13	みんなの広場：自分を叱ることができるか＝フリーター・大六野卑朋・２４	×
14	社告：アートフェス 現代美術を身近に―来月１４日から、兵庫・尼崎 【大阪】	×
15	社告：農コン入賞者の生産物を産直で―スポンデプラザ大阪 【大阪】	×
16	社告：西日本選抜学生相撲大会 技と力が激突―来月８日・堺市 【大阪】	×
17	社告：「ごんぎつねの世界」展 南吉の愛と悲しみ―堺市立東文化会館 【大阪】	×
18	ファイル：深谷元通産相、引退表明	×
19	ファイル：オスプレイ配備、撤回要求	×
20	社告：世界遺産ヴェネツィア展 華麗な歴史、魅惑の芸術 前売り券発売中―京都文化博物館 【大阪】	×
21	首相日々：２２日	×
22	社説：原子力基本法 「安全保障目的」は不要	×
23	自民党総裁選：林氏、改めて出馬意欲―名古屋で講演	×
24	松田喬和の首相番日誌：小沢元代表の純化路線	×
25	社説：延長国会 与野党は７９日を生かせ	×
26	ファイル：官房長官が体調不良	×
27	税と社会保障：一体改革審議 民自公、かばい合い答弁 玉虫色の合意、取り繕う	×
28	首相日々：２１日	×
29	小沢・民主元代表：「妻の手紙」出回る 大量コピー、民主議員らに郵送	×
30	東日本大震災：福島第１原発事故 米の放射線情報、地元の説明要請―復興相	×
	ベースラインにより抽出された特徴語	
	者、高台、サイクル、大手、事故、稼働、住宅、計画、性、中、経済、負担、燃料、核、移転、事業、最終、必要、避難、帰還、エネルギー、除、確保、電気、こと、年、貯蔵、被害、基本、多く、的、停止、自治体、全国、地域、住民、原発、価格、汚染、よう、費、促進、発送、先月、問題、以外、可能、処理、処分、設置、声、解除、安全、仮設、電、国、送電、県、中間、電力、対策、化、供給、被災、地、政府、再生、制度、安定、福島、道筋、依存、料金、復興、送配、温暖、大量、使用、年間、分離、網、立地、課題、発電、量、新規、染、難題、責任、	

表 6 特徴語抽出：正解例 1

正解見出し
ヤングフェスティバル：千葉・～ふれあい ・絆～地域の人たちとの交流 / 千葉
見出し語
ヤング, フェスティバル, 千葉, 絆, 地域, 人, たち, 交流
抽出された特徴語
千葉, 県, 交流, 青少年, 地域, グルメ, コーナー, 町, 人, ヤング, フェスティバル, 昨年, 協会, 企画, 高, たち, 入場,

表 7 特徴語抽出：正解例 2

正解見出し
東日本大震災：福島第 1 原発事故 「避難拡大指示、頭越し」 福島県知事、政府批判ー国会事故調
見出し語
東日本, 大震災, 福島, 原発, 事故, 避難, 拡大, 指示, 頭, 越し, 県知事, 政府, 批判, 国会, 事故調,
抽出された特徴語
情報, 省, 事故, 会, 確認, 指示, 計画, 産, 性, 政府, チェック, 保安, 経, 避難, 院, 原子力, こと, 側, 委員, 原発, 国会, 委, 事故調, キロ, 安全, 質疑, 評価,

表 8 特徴語：不正解例 1

正解見出し
視聴室：湯けむりスナイパー・お正月 S P
見出し語
視聴, 室, 湯けむり, スナイパー, お正月, S P
抽出された特徴語
屋, さん, ワケ

表 9 特徴語抽出：不正解例 2

正解見出し
シンポジウム：「作家の越境」テーマに日本大で リービ英雄さんが基調講演
見出し語
シンポジウム, 作家, 越境, テーマ, 日本, 大, リービ, 英雄, さん, 基調, 講演
抽出された特徴語
者, 日本語, 教授, 作家, 賞, 受賞, シンポジウム

表 10 抽出できた特徴語の数

頻度	特徴語の数
16	0
15	1
14	2
13	3
12	0
11	9
10	22
9	26
8	34
7	31
6	73
5	182
4	420
3	7162
2	16721
合計	24686

表 11 抽出できたパターンの数

頻度	パターンの数
8	0
7	2
6	2
5	14
4	32
3	532
2	14846
合計	15428