各種 Web 新聞記事データからの話題語抽出

櫻田　殊久†　馬　青†　村田　真樹‡
†龍谷大学大学院理工学研究科
‡情報通信研究機構

１．はじめに

話題語とは人々の間で注目され、世間を賑わせた言葉のことである。話題語抽出は、世の中に洪水のように氾濫している多種多様な情報に対する究極の情報集約手段として捉えることができる。本研究が目指す話題語は朝日新聞社が開催していた「ワード・オブ・ザ・イヤー」[1]のような、その年人々の間でよく話されて、ニュースや新聞などでよく取り上げられる事柄などを象徴した言葉である。本稿ではこのような年間話題語を、Web 新聞記事データの自動収集を行う「HiMakeDoc」[2]というフリーソフトを利用し、収集した１年間分の各種新聞記事（2007 年の場合、計 77 サイト、192,000 記事）から抽出する手法を提案する。

２節で話題語の関連研究について、３節で話題語の候補となる言葉について述べた後、４節と５節で話題語の候補として複合名詞の抽出と未知語の復元について述べる。６節で話題語候補に対する順位付け（すなわち話題語抽出）の手法を説明する。７節で Web 新聞記事を対象とした話題語抽出の実験結果を示すとともに考察を行う。最後に８節でまとめを述べる。

２．関連研究

話題語抽出の研究は今までに複数提案されている。佐藤ら[3]は新聞記事に対し、文書クラスタリングを用いて同一話題を集約した後、類似文書の量によって文書の持つ話題性と文書内の語句の話題性を総合して話題語の抽出を行った。しかし、この手法は半月ごとの数千程度の記事を対象とした話題語の抽出を対象としており、本稿のような１年間を対象とした大規模な記事データへの適用は処理時間が莫大に増加するため困難である。また、手法自体が話題語の順位付けが主題となっており、話題語の候補となる言葉の抽出方法についてはあまり着目されていなかった。

一方、新聞記事ではなく blog から話題語を抽出する手法も提案されている[4]。これは本研究が目指している話題語とは異なり、ブログの書き手同士の興味の関連度を求めることにより興味が同じ人々の間で共通して用いられる言葉を探すことを目的としている。また、研究[3]と同様、手法自体は話題語の順位付けが主題となっている。このような話題語の順位付けを主題とする研究は他にもある[5]。

本研究は、上記の問題点を踏まえ、話題語候補の抽出と順位付け（話題語抽出）の新しい手法を提案する。提案手法は年間を通した言葉の頻度（出現回数）の変化に着目したものであり、記事数が数十万に上る大規模データにも簡単に適応できる。

３．話題語の候補となる言葉

例年の「ワード・オブ・ザ・イヤー」に選ばれた話題語のタイプを調べると、その数の多い順に(1)単名詞・複合名詞（人名・固有名詞を含む）、(2)フレーズ、(3)未知語、(4)その他、になっている傾向が見られる。本稿では(2)を今後の課題とし、(1)と(3)から話題語の候補を抽出する。ただし、単名詞の抽出が簡単なため、候補抽出については複合名詞と未知語について述べる。

４．複合名詞の抽出

たとえば「郵政民営化」が話題語であればそれを構成する「郵政」と「民営化」の単語は文中において他の名詞と一緒に使用されることがあまりない。すなわち、話題語の候補となる複合名詞を構成する単語同士は強い結びつきがあると言える。その結びつきを考慮し複合名詞を取り出す手法として同期性の計算に基づくものがある[6]。一方、その結びつきを測るには共起頻度を使うことも当然考えられる。本研究では、従来の同期性手法を改善し上で、共起頻度に基づく手法との比較を行った。その結果、共起頻度に基づく手法の方が優れていることがわかり、複合名詞の抽出に用いることにした。

4.1 同期性手法

これは、複合名詞を構成する単語の各々の単語頻度は時系列推移で同期しているという仮定に基づく手法である。しかし、従来手法[6]は、同期性の指標となるスコアを計算した際にスコアの大きさが単語の頻度に影響されるため、複合名詞を特定する閾値の設定が困難という問題がある。また、従来手法では複合名詞を構成する単名詞から

重要語を1つ選び、その重要語と他の単語の同期性を計算している。しかし、重要語の決定はそう簡単ではないし、こうする必要性もあまり感じない。そこで本研究では以下の改良を行った。

ここで、単語のある期間の出現頻度の差分に関するベクトルを導入する。時刻t_kの単語w_iの頻度を$f(w_i,t_k)$、時刻t_{k+1}の時の頻度を$f(w_i,t_{k+1})$と表すと、単語w_iの時刻t_kからt_{k+1}までの出現頻度の差分ベクトルを(1)式で求める。

$$V_t(w_i,t_k,t_{k+1}) = (f(w_i,t_{k+1}),t_{k+1}) - (f(w_i,t_k),t_k) = (f(w_i,t_{k+1}) - f(w_i,t_k),t_{k+1} - t_k) \quad (1)$$

差分ベクトルを用いて単語w_iとw_jの時刻t_kからt_{k+1}間の頻度の同期余弦尺度を用いて計算すると(2)式になる。

$$\cos_t(w_i,w_j,t_k) = \frac{V_t(w_i,t_k,t_{k+1}) \cdot V_t(w_j,t_k,t_{k+1})}{\sqrt{(V_t(w_i,t_k,t_{k+1}))^2} \times \sqrt{(V_t(w_j,t_k,t_{k+1}))^2}} \quad (2)$$

これを同期性を計算したい期間(t_0からt_{n-1})に拡張し、足し合わせることで同期性の時間推移の相違度を計算することできる。

$$\cos_T(w_i,w_j) = \sum_{k=0}^{n-1} \cos_t(w_i,w_j,t_k) \quad (3)$$

さらに$\cos_T(w_i,w_j)$を、計算した区間分で割り、平均を計算する。

$$\cos_{mean}(w_i,w_j) = \frac{\cos_T(w_i,w_j)}{n-1} \quad (4)$$

$\cos_{mean}(w_i,w_j)$は必ずあらゆる単語の組み合わせに関わらず-1から1の間の値を取る。したがって、複合名詞を抽出する閾値を設定することが可能となる。この時、単語w_iとw_jは隣接する単語同士であり、重要語をあらかじめ決めておく必要はない。

4.2 共起頻度手法

単語間の結びつきは単語間の共起で見ることができる。bigramによる共起頻度を用いた複合名詞の抽出手順は以下の通りである。

①記事を形態素解析し、連続して出現する名詞列(a, b, c, d, …) を抽出する。

②①で抽出した名詞列の先頭からab, bc, cd, …の共起頻度を計算する

③ 閾値を用いて共起頻度が閾値以下の場合は結合せず抽出する。すなわち、cdの共起頻度が閾値以下なら、dを結合せずabcを複合名詞とする。このとき、部分列(abとbc)も複合名詞とする。

4.3 同期性手法 vs. 共起頻度手法

両手法の優劣を比較するために、過去(2005年と2006年)のWeb新聞記事データを比較実験に用いた。同期性手法においては各時刻($t_0, t_1, \cdots, t_{n-1}$)を半月ごととし、閾値を0から0.1刻みで変化させた。共起頻度手法においては閾値を2, 5, 10, 20, 50, 100, 500, 1000と設定した。実験方法としてそれぞれの年から話題語になりそうな複合名詞を100個人手で取り出してリストを作成し、閾値の変化によってリストにある複合名詞が得られる個数を計算した。ただし、作成したリストが100単語しかなく、そのリストにabcが入っていてabcxyz…が入っていない場合が考えられる。abcxyz…が複合名詞であればabcを複合名詞とみなさないようにする必要があるため、この比較実験に限って両手法の最後に次の手順を追加した。

abcを含む表現abcxyz…を記事から取り出し、再度同じ閾値を用いてabcxyz…が複合名詞とするかを確認する。abcxyz…を複合名詞として判断されれば、その部分列のabcを複合名詞とみなさない。ただし、x,y,z,…は名詞の場合に限る。

表1 同期性手法で抽出できた複合名詞の数

閾値	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2006年	21	21	47	46	40	30	21	16	10	7
2005年	61	65	63	63	57	46	35	24	17	13

表2 共起頻度手法で抽出できた複合名詞の数

閾値	2	5	10	20	50	100	500	1000
2006年	71	84	85	89	91	91	48	25
2005年	80	94	97	94	91	90	61	38

実験結果を表1と表2に示す。表より、複合名詞の抽出には共起頻度手法の方が優れていることがわかった。また、共起頻度手法において、閾値100を境にして急激に候補の数が落ちていることと、安定的に閾値が2-100の間で結果に差が出ないことを考慮し、閾値100程度で抽出を行うのがよいこともわかった。

5．未知語の抽出

形態素解析結果に未知語として判断される形態素は人名や地名の一部分と思われる場合が多く見

られ、復元処理が必要となる。未知語を復元する手法に[7]がある。

この手法では、未知語を復元するためには未知語と判断された文字とその前後の文字列が必要となるので未知語を含んだ文節を係り受け解析ツール南瓜にて切り出す。ただし、未知語が「ひらがな」の場合は、まず以下のルールにて文節を拡張する。

① 未知語が「ひらがな」の場合、前後の文字に「ひらがな」が続く限り、文節を拡張する。
② 違う文字種の文字にぶつかったら違う文字種の文字を含む文節までを切り出す文節拡張の範囲と定める。

次に切り出した文節と文書中にある同じ未知語を含んだ文と未知語を中心に前後比較し、一致する前後の部分を取り出す。しかし、上記手法では「フィッシング詐欺や」の「や」、または「セルティック・中村」の「・中村」といったような余計なものがついている場合が見られる。このような問題を解決するために、以下の二つルールを追加した。

③未知語がカタカナ語で「・」を含み、「・」を境にして文字種が変わる場合は「・」を境にして未知語を含まない部分を削除する（「・」も同時に削除する）。

④未知語を復元すると人名になることがある。候補の最後尾に「さん」がついている場合、「さん」を削除する。

以上の四つのルールを用いて未知語候補を得る。次にそれらの頻度を用いて未知語の抽出を行う。ただし、極端に頻度が低い場合を除けば頻度が高い候補ほど文字列が短くなる傾向がある。そこで閾値を設けて閾値以上で閾値に一番近い頻度の候補を抽出し、未知語とする。このような未知語は実際、閾値以上でもっとも長い文字列のものになる。

過去の１年分の新聞データ(2005 年)を用いた未知語抽出の予備実験を行った。その結果、上記手法は有効であることが確認できた。また、閾値を 20 に設定するのがよいこともわかった。

６．話題語抽出

抽出した複合名詞、未知語、そして単名詞を話題語の候補とし、それらを順位付けすることにより話題語を抽出する。話題語の抽出には、年間を通した言葉の頻度の変化に着目した手法を用いる。話題語の多くは話題として人々に注目されると急激に頻度が上昇する特徴がある。ここで一般語「夢」と話題語「ハンカチ王子」の半月毎の頻度のグラフを示す（図 1）。このグラフを見ると「ハンカチ王子」は８月上旬から８月下旬にかけて頻度が急激に上昇していることがわかる。ここで話題語の各候補の頻度上昇の度合いをその候補の話題度ととらえ、その計算にいろいろな方法が考えられる。それらの方法に対し、過去の新聞データ（2005 年、2006 年）を用いた比較実験を行うことにより、以下の計算式を採用した。

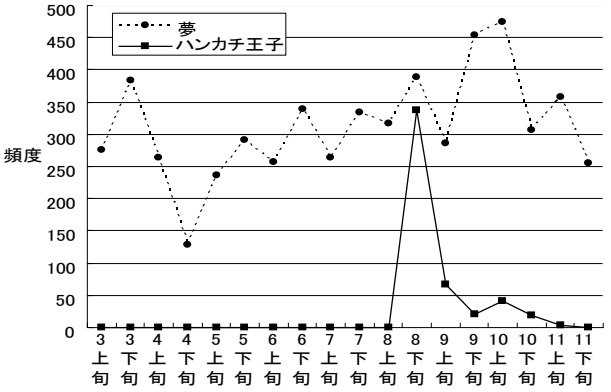

図１ 一般語「夢」と話題語「ハンカチ王子」の頻度変化

$$TD = \frac{\Delta f}{\bar{f}_l} \tag{5}$$

ただし、TD は話題度、Δf は調査する年における頻度上昇の度合い、$\bar{f}_l$ はその前の年の半月ごとの平均頻度である。Δf は以下のように求める。

$$\Delta f = f_h - f_s \tag{6}$$

ただし、f_h は頻度が最も高い半月の頻度であり、f_s は頻度が上昇し始める半月の頻度である（図１の場合、「夢」と「ハンカチ王子」の f_s はそれぞれ９月上旬と８月上旬となる）。また、$\bar{f}_l$ が 0 または１以下の場合は１として計算する。これは 0 の場合は(5)式の計算ができず、１以下の場合 TD の値を必要以上に増大させてしまうからである。なお、頻度のもっとも高い月が１月上旬の場合は前年度のデータを用いる。

７．評価実験

話題語抽出の評価実験には 2007 年（1 月から 11 月）の新聞データを用いた。それは 77 サイトからの計 192,000 件の記事から構成される。（過去データを用いた予備実験により）複合名詞の抽出には共起頻度手法を用い、その閾値を 100 に設定した。未知語の抽出にその閾値を 20 に設定した。

表 3　抽出した 2007 年の話題語

	話題語	TD	未知語	TD
1位	中越沖地震	3356	IXI	296
2位	都知事選	715.44	三岐鉄道・阿下喜駅	274
3位	柏崎市	645	セントラルタワーズ	260
4位	浅野氏	635	ツール・ド・ロマンディ	258
5位	新潟県 中越沖地震	614	アニータ	239
6位	佐賀北	606	PKK	205
7位	ミートホープ	561	パイル	203.87
8位	食肉偽装	469	ボルヘッティ	192
9位	カーリング協会 競技委員長	448	エナン	184
10位	特待制度	423	クラシエ	183
11位	リンナイ	392	LABI	178
12位	伊藤市長	391	キベト	178
13位	西武現金供与	388	ギゾ	177
14位	意見聴取会	381	セイヨウオオマルハナバチ	174
15位	阿部氏	379.077	ダン・ラザー氏	158
16位	福知山線脱線	365	ダビンチ	150
17位	アパホテル	360	ジョブカード	138
18位	久間防衛相辞任	356	GWG	136
19位	元専務	351	リエティ	135
20位	記録都市対抗野球	350	アイメイト	134

話題語の候補の抽出において、単名詞 53,426 個、複合名詞 121,587 個、未知語 51,837 個を抽出した。それらの候補に対する順位付けの結果(上位 20 位の話題語）を表 3 の左列に示す。上位 10 位以内の言葉に注目すると「中越沖地震」と「柏崎市」と「新潟県中越沖地震」は 2007 年 7 月の新潟県中越沖地震関連の言葉である。「都知事選」と「浅野氏」は 2007 年 4 月の東京都知事選の関連の言葉で、「佐賀北」と「特待生制度」は高校野球関連、「ミートホープ」と「食肉偽装」は 2007 年の 6 月のニュースである。これらは総じて 2007 年の話題語と言えよう。表 3 の右列に未知語の上位 20 位を載せている。ただし、未知語の候補は話題語上位の 41 位以内には入っていない。これは 2007 年に話題になった言葉の中に未知語は少ないためと考える。しかし、未知語の復元の効果は確認できた。例えば未知語の 2 位の「三岐鉄道・阿下喜駅」や 4 位の「ツール・ド・ロマンディ」などは復元を行わなければ得られない言葉である。また、未知語の復元手法を改良した効果として、たとえば、「先発・ローウェン」の代わりに「ローウェン」、「キム・ヤンヒャンさん」の代わりに「キム・ヤンヒャン」が得られた。

また、話題語抽出（話題語候補への順位付け）の提案手法と従来手法[3]との比較実験も行った。手法[3]は処理時間がかかるため、記事数を実験の行える範囲に減らした（ジャンルごとにおよそ 120～1000 記事）ものである。手法[3]については、半月毎ごとに抽出した上位 5 位の言葉に含まれる話題語の数と計算時間を測定した。その結果、抽出した全 110 個の言葉に含まれる話題語は 14 個であり、12%の正解率であった。一方、提案手法について上記手法[3]の実験に合わせて抽出した上位 110 個の言葉に含まれる話題語の数を測定すると、35 個と、手法[3]よりはるかに高い 32%の正解率であった。また、実験時間（候補への順位付け）を比較すると。手法[3]は 27,944 秒(およそ 7.7 時間)で、提案手法は 1,926 秒(およそ 32 分)であった（プログラムは Perl を用いた）。

8．まとめ

本研究は大規模 Web 新聞記事データから年間の話題語候補（単名詞、複合名詞、未知語）の抽出とそれらの順位付けによる話題語抽出の新しい手法を提案した。提案した話題語抽出手法は年間を通した言葉の頻度の変化に着目したものであり、記事数が数十万に上る大規模なデータにも簡単に適用できる。計算機実験の結果、比較的に精度よく年間話題語を取り出すことができた。また、従来手法より抽出精度と計算時間の両面において勝っていることが確認できた。今後は抽出精度のさらなる向上を図るとともに、話題語となりうる他のタイプの言葉、すなわち、「形容詞＋名詞」、「名詞＋の＋名詞」、「フレーズ」からの話題語抽出へ拡張する予定である。

参考文献

[1] 朝日新聞社主催「ワード・オブ・ザ・イヤー」 http://opendoors.asahi.com/index.shtml
[2] HiMakeDoc http://homepage3.nifty.com/hiro_ish/palm/himakedoc/hmdoc.htm
[3] 佐藤、川島、佐々木、大久保：文書の類似度と新鮮度に基づく話題語抽出、2005-NL-165(5)
[4] 関口、佐藤、川島、奥田、奥雅：blog ページ集合に対する話題語抽出手法、2005-NL-170(5)
[5]佐藤、坂井、川島 、奥田：単語出現の意外性に基づく話題性評価方法、2007-NL-181
[6] 村上、渡辺：時系列情報用いた複合語キーワード、2006-NL-172(1)
[7] 石原、 山田、 松本、 池田： 1 文字未知語からの未知語候補の復元、NLP2006