キーワードの活性度の変化を用いたテキスト中の単語と話題との対応付け

高橋 雅仁　吉村 賢治　首藤 公昭
福岡大学 工学部

1 はじめに

筆者らは，意味ネットワークにおける単語の活性状態を疑似的に捉えることが可能な単語の共起関係を基に構築した文脈情報を用いて，テキストセグメンテーション [1] やテキストからの話題構造抽出 [2] を行う手法を提案している．本稿では，上記の話題構造抽出手法を用いて抽出したテキストの話題構造中の各話題とテキスト中の各単語との意味的な関連性の強さを定量的に求める手法を提案する．

本手法では，単文中の名詞と動詞間の共起関係に基づいて入力テキストに対する文脈情報を動的に構築し，得られた文脈情報中の特に話題の変化に応じて活性度が変化しやすい単語群（以降,「キーワード」と呼ぶ）に着目して，それらの活性度の変化を調べ，テキストの話題構造を抽出する．抽出した各話題は，キーワードの部分集合として表現される．この部分集合に含まれる各キーワードが入力テキスト中のどの単語によって活性化されたかを文脈情報を参照して調べることにより，テキスト中の各単語と話題構造中の各話題との意味的な関連性の強さを定量的に求めることができる．

以下，2章で文脈情報の生成，3章で話題構造抽出，4章でテキスト中の単語と話題との関連付けのための各アルゴリズムを示す．5章では本手法によるテキスト中の単語と話題との関連付けの例を示し，考察を加える．最後に6章でまとめを行う．

2 文脈情報の生成

単文内での名詞と動詞の共起情報を2項組 (N, S_V), (V, S_N) で表す．ここで，Nは名詞，Vは動詞，S_VはNを格要素としてとる動詞とその出現頻度の対の集合，S_NはVの格要素となる名詞とその出現頻度の対の集合である．たとえば，名詞「雪」に対する共起情報は，$(雪, \{(降る, m_1), (積もる, m_2), (警戒する, m_3), (滑べる, m_4), \cdots\})$，また，動詞「降る」に対する共起情報は，$(降る, \{(雨, n_1), (雪, n_2), (北国, n_3), (天, n_4), \cdots\})$ となる．文脈情報は，名詞Nとその累積刺激kの対の集合$\Sigma = \{(N_1, k_1), (N_2, k_2), \cdots, (N_l, k_l)\}$によって表す．集合$\Sigma$は，テキストの先頭から単語（名詞に限る）を順次読み込み，各単語に対して以下の文脈情報の更新処理を施し生成する．なお，集合Σの初期値は空集合とする．

1. 読み込んだ名詞Nについて，共起辞書より共起情報(N, S_V)を取り出す．
2. S_V中のすべての動詞$V_i (i = 1, 2, \cdots, m)$について，共起辞書より共起情報$(V_i, S_{N_i})$を取り出す．
3. S_{N_i}中のすべての名詞$N_j (j = 1, 2, \cdots, n_i)$を集合$\Sigma$に加える．ただし，名詞$N_j$の累積刺激$k_j$は，$N \cdot V_i$間の共起の出現頻度$a_i (i = 1, 2, \cdots, m)$と$V_i \cdot N_j$間の共起の出現頻度$b_j (j = 1, 2, \cdots, n_i)$の関数$f(a_i, b_j)$（たとえば，$k_j = a_i \cdot b_j$）で与える．なお，同じ名詞が既に集合$\Sigma$に存在する場合は，その名詞の累積刺激の加算のみを行う．

3 テキストからの話題構造抽出

2章の処理で得た集合Σの要素となっている単語の累積刺激値の時系列的な変化に着目すると，その変化率（活性度）がテキストの話題境界で大きく変化するような単語は，テキストの話題との意味的な関連性が高い単語だと言える．本手法では，このような話題の変化によく反応する単語をキーワードとして予め選定しておき，入力テキストに対して動的に構築される集合Σの中のキーワードの活性度の変化に着目してテキストの話題構造を抽出する．

なお，意味ネットワークを用いた従来のテキストセグメンテーション手法 [3][4] では，意味ネットワーク上で活性化された入力テキスト中の単語の活性度に着目して話題境界を求めている．一方，本手法では，入力テキスト中の単語とは独立に設定されたキーワード集合の活性度の変化に着目している．このため，本手法は，従来手法と比較して，言語知識（本研究においては，名詞と動詞間の共起情報）の希薄性の問題に対する頑強性が高いという特長をもつ [1]．

以下に話題構造抽出の処理手順を記す．

1. 入力テキストに対して集合Σの生成を行う．このとき，入力テキストの先頭から見てt番目の名詞$N_t(t = 1.2.\cdots.l)$に対する文脈情報の更新処理が終了する毎に，予め定めたキーワード$X_i(i = 1.2.\cdots,m)$に対する文脈情報の累積刺激$k_{i_t}(i = 1.2.\cdots,m.t = 1.2.\cdots.l)$の値を集合Σから読み出し，記録していく．

2. 各キーワードX_iに対して，累積刺激k_{i_t}の時系列的な変化を近似する関数$g_i(t)$を求め，さらに，その2階微分を行って関数$g_i''(t)$を得る．

3. 各キーワードX_iに対して，関数$g_i''(t)$の極大値の発生位置(話題の開始位置)を始点とし，極小値の発生位置(話題の終了位置)を終点とする線分L_iを作成する．なお，線分L_iは，上記の極大値および極小値の絶対値の和によって求められる線幅をもつ．

4. 線分L_iをその長さが長いものから順に以下の条件を満たすように積み重ねて山型$M_j(j = 1.2.\cdots.n)$を作る．1つの山型を作り終えたら，残った線分を用いてさらに山型を作る処理を繰り返す．

   - 積み重ねる線分の始点および終点は，その土台となる直下の線分上に位置する．
   - 積み重ねる線分と直下の線分の長さの差は，ある定められた値よりも小さい．

5. 得られた山型M_jからテキストの話題構造を抽出する．

なお，キーワード$X_i(i = 1.2.\cdots,m)$は，テキストの話題境界位置を示すタグを付与したキーワード学習用テキストを使って，話題の変化によく反応する名詞，すなわち，話題境界の近くで文脈情報の累積刺激の2階微分の極値が発生しやすい名詞を選別するキーワード学習によって求める．キーワード学習は話題構造抽出処理に先立ち1回だけ実行される．

4 テキスト中の単語と話題との関連付け

テキスト中の単語と話題との関連付けのために，話題構造抽出処理に引続き，以下の処理を実行する．

1. 話題構造抽出の処理で得た各山型$M_j(j = 1.2.\cdots.n)$を構成するキーワードを$X_p(p = 1.2.\cdots.r_j)$とする．

2. 話題構造抽出のステップ1の処理で記録した各キーワードに対する文脈情報の累積刺激値を参照して，入力テキスト中の名詞$N_t(t = 1.2.\cdots.l)$と山型$M_j(j = 1.2.\cdots.n)$との関連度Rを以下の数式により求める．

$$R = \sum_{p=1}^{r_j}(k_{p_t} - k_{p_{(t-1)}})$$

   ここで，k_{p_t}は，名詞N_tの入力直後のキーワードX_pに対する文脈情報の累積刺激値を意味する．また，$k_{p_0} = 0$とする．

なお，話題構造抽出の処理で得た各山型M_jの中には，複数の異る話題が含まれている可能性がある．そこで，各山型M_jを構成するキーワード$X_p(p = 1,2,\cdots,r_j)$からなる集合をシソーラスなどを用いて同じ概念に属するキーワードのグループ毎に類別すれば，これらの概念と入力テキスト中の名詞との関連度Rを求めることもできる．

5 実験

5.1 共起辞書の作成

まず,「EDR電子化辞書日本語共起辞書第2版[5]」から8種類の格助詞("が", "を", "に", "へ", "と", "から", "より", "で")をとる名詞と動詞間の共起レコードを抽出した(217,407レコード)．次に，これらの共起レコードを用いて名詞を検索キーとする名詞と動詞間の共起レコードを格納した共起辞書と，動詞を検索キーとする動詞と名詞間の共起レコードを格納した共起辞書を作成した．

5.2 キーワードの学習

「日本経済新聞CD－ROM版(90年～94年，計5年分)[6]」からコラム「春秋」の記事データ1,210日分(総文数：22,169，総段落境界数：3,889)を抽出し，キーワード学習を行った．キーワード学習では，記事データの形式段落を意味段落とみなし，共起辞書に含まれる名詞19,781語の中から，記事データの段落境界で活性度が比較的よく変化した3,162語の名詞をキーワードとして選択した．なお，上記のテキストファイルからの処理対象とする名詞類[1]の抽出には,「日本語形態素解析システムJUMAN version 3.5[7]」を使用した．

[1] JUMANの品詞分類のうち，'普通名詞'，'サ変名詞'，'固有名詞'，'時相名詞'，'名詞性名詞助数辞'，'名詞性名詞接尾辞'，'名詞接頭辞'，'カタカナ'の8種を処理対象とした．

5.3 実行結果

ここで，本方式によるテキスト中の単語と話題との関連付けの具体例を示す．図1は，日本経済新聞のコラム「春秋」(1994 年 2 月 28 日付) 中の連続する 2 段落からなるテキストである．なお，このテキストはキーワード学習では使用してしない．このテキストを入力テキストとし，3,162 語のキーワードを用いて各キーワードへの累積刺激の 2 階微分の極値の総和を求めたグラフを図2に示す．図2を見ると，段落の境界付近で極大値の総和が突出しており，話題境界 (段落境界) が捉えられていることがわかる．図 3には，入力テキストに対して，話題構造抽出処理を行って得た話題構造を表す山型を示す．なお，図3においては，山型の高さが，もっとも高い山型 (話題3) の高さの20%未満となった山型は出力していない．図3より，この入力テキストは，テキスト全体に関連する話題 (話題1)，テキストの前半の話題 (話題2)，テキストの後半の話題 (話題 3 と話題 4) によって構成されていることがわかる．なお，これらの話題を表す山型を構成するキーワードの数は，話題 1 で 144 語，話題 2 で 275 語，話題 3 で 676 語，話題 4 で 369 語であった．各話題の内容は，それぞれの話題に含まれるキーワードの集合によって表現される．そこで，これらのキーワード集合をシソーラスを用いて同じ概念で類別し，各話題のタイトルおよびサブタイトルを与えることを別途検討中である．

表1には，入力テキスト中の各単語 (名詞類) と話題構造抽出処理で得た各話題との関連度を記している．なお，表1中の入力単語の中には，形態素解析処理の誤りによって得た単語が含まれている．また，表1において，特に，抽出した話題との関連度が高いテキスト中の単語 (各話題中で関連度がもっとも高く，かつ，その値が 0.015 以上となった単語) には，関連度の数値に★印を付与している．テキスト中の★印の付いた単語の分布は，ほぼ，話題を表す山型の出現位置と一致していることがわかる．なお，表1において，テキスト中の固有名詞 (「ルツコイ」と「細川」) の関連度が 0 となっているが，これは，名詞，動詞間の共起辞書に固有名詞に関する共起情報が不足しているためである．固有名詞も話題を表す重要な単語と考えられるので，固有名詞の扱いについては，今後検討が必要と思われる．

6 おわりに

意味ネットワークにおける単語の活性状態を疑似的に捉えることが可能な，単語の共起関係から求めた

このところ，エリツィン大統領率いる改革派の旗色は悪い。
しかし押せ押せムードの保守派にも悩みはあった。
議会下院が決めた恩赦の対象者は迷った。
無罪の者が恩赦を受けてはおかしい。
へたをすれば有罪を認めたことになる。
釈放されたルツコイ元副大統領らは、有罪か無罪かはっきりしない起訴取り下げ措置に期待しているのか。

---------- 段落境界 ----------

週末の七カ国蔵相・中央銀行総裁会議（G７）は、型通り対ロシア支援問題を話し合った。
数百億ドルの支援額を決めたかつての熱気はどこへやらだった。
一方、羽田外相の三月モスクワ訪問がほぼ決まったものの、細川内閣の改造騒ぎで外相交代説が出ている。
訪問の日程は詰めにくい。
「“変調”は何もわれわれの側だけでない」というつぶやきが、ロシアの方から聞こえてくるようだ。

図 1: 入力テキストの例

文脈情報を用いて，テキストから抽出した話題構造中の各話題とテキスト中の各単語との意味的な関連性の強さを定量的に求める手法を提案した．今後は，話題構造中の各話題を表すキーワード集合の分析作業などを行い，本方式のさらなる改善を図る予定である．また，本手法の情報検索や要約への応用を図りたい．

謝辞 本研究遂行のため，「日本語形態素解析システム JUMAN」，「EDR 日本語共起辞書」，「日本経済新聞 CD-ROM 版」を利用させて頂きました．それぞれの開発・提供に係わる方々に感謝致します．また，実験システムの作成などでご協力を頂いた (株) エクシーズの森澤慎一郎氏，福岡大学工学部電子情報工学科知能工学研究室の諸氏に感謝致します．

参考文献

[1] 高橋 雅仁，森澤 慎一郎，吉村 賢治，首藤 公昭 ：キーワードの活性度の変化を用いたテキストセグメンテーション，2000 年情報学シンポジウム講演論文集，pp.145-152 (2000).

[2] 高橋 雅仁，吉村 賢治，首藤 公昭 ：キーワードの活性度の変化を用いたテキストからの話題構造抽出，第 52 回電気関係学会九州支部連合大会講演論文集，p.674 (1999).

[3] 小嶋 秀樹，古郡 廷治 ：単語の結束性にもとづいてテキストを場面に分割する試み，情報処理学会研究報告 93-NL-95，pp.49-56 (1993).

[4] Olivier Ferret ：How to thematically segment texts by using lexical cohesion?，Proc. COLING-ACL '98，pp.1481-1483 (1998).

[5] (株) 日本電子化辞書研究所　共起辞書 (第 2 版)(TR-043)，(1994).

[6] 日本経済新聞社編　日本経済新聞ＣＤ－ＲＯＭ版 (90 年～94 年)，(1995).

[7] 黒橋 禎夫，長尾 真　日本語形態素解析システム JUMAN version 3.5，(1998).

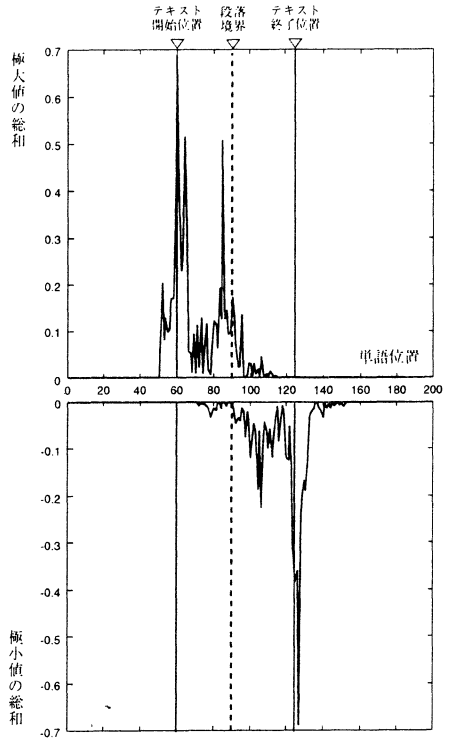

図 2: 入力テキストに対する累積刺激の 2 階微分の極値の総和

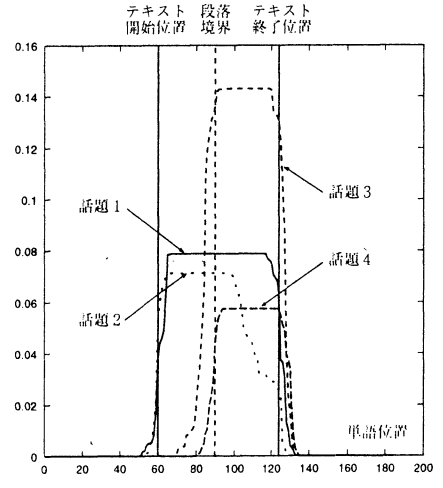

図 3: 入力テキストに対する話題構造

表 1: テキスト中の単語と話題の関連度

No.	入力単語	話題 1	話題 2	話題 3	話題 4
1	大統領	0.0086	0.0101	0.0120	0.0016
2	改革	0.0137	0.0050	0.0076	0.0023
3	派	0.0117	0.0092	0.0039	0.0015
4	旗色	0.0000	0.0000	0.0000	0.0000
5	押せ押せ	0.0000	0.0000	0.0000	0.0000
6	ムード	★0.0157	0.0109	0.0105	0.0077
7	保守	0.0135	0.0045	0.0063	0.0022
8	派	0.0117	0.0092	0.0039	0.0015
9	悩み	0.0150	★0.0306	0.0233	0.0006
10	議会	0.0109	0.0087	0.0098	0.0015
11	下院	0.0027	0.0046	0.0020	0.0000
12	恩赦	★0.0204	0.0168	0.0072	0.0003
13	対象	★0.0382	0.0094	0.0062	0.0010
14	者	0.0063	★0.0170	0.0135	0.0022
15	無罪	0.0105	★0.0195	0.0006	0.0000
16	者	0.0045	★0.0161	0.0081	0.0011
17	恩赦	★0.0204	0.0168	0.0072	0.0003
18	へた	0.0000	0.0015	0.0000	0.0000
19	有罪	0.0115	0.0046	0.0015	0.0000
20	釈放	0.0125	★0.0247	0.0051	0.0000
21	ルツコイ	0.0000	0.0000	0.0000	0.0000
22	元	0.0014	0.0018	0.0018	0.0002
23	副	0.0000	0.0000	0.0000	0.0000
24	大統領	0.0086	0.0101	0.0120	0.0016
25	ら	0.0000	0.0000	0.0000	0.0000
26	有罪	0.0115	0.0046	0.0015	0.0000
27	無罪	0.0105	★0.0195	0.0006	0.0000
28	起訴	0.0016	0.0003	0.0003	0.0000
29	取り下げ	0.0017	0.0056	0.0000	0.0000
30	措置	0.0254	★0.0388	0.0095	0.0003
31	期待	0.0090	0.0161	★0.0457	0.0006
32	週末	★0.0168	0.0091	0.0122	0.0000
33	カ国	0.0012	0.0014	0.0008	0.0012
34	蔵相	0.0047	0.0135	★0.0231	0.0012
35	中央	0.0109	0.0132	★0.0397	0.0067
36	銀行	0.0166	0.0104	★0.0201	0.0039
37	総裁	0.0154	0.0107	★0.0205	0.0043
38	会議	★0.0236	0.0107	0.0145	0.0031
39	型通り	0.0000	0.0000	0.0000	0.0000
40	対	0.0000	0.0000	0.0000	0.0000
41	支援	0.0077	0.0061	★0.0285	0.0082
42	問題	0.0184	0.0066	★0.0240	0.0029
43	ドル	0.0064	0.0042	0.0137	0.0123
44	支援	0.0077	0.0061	★0.0285	0.0082
45	額	★0.0260	0.0058	0.0186	0.0072
46	熱気	0.0019	0.0031	★0.0153	0.0450
47	へや	0.0000	0.0000	0.0000	0.0000
48	ら	0.0000	0.0000	0.0000	0.0000
49	外相	0.0028	0.0032	★0.0263	0.0029
50	月	0.0140	0.0043	0.0133	0.0027
51	訪問	★0.0202	0.0055	0.0200	0.0155
52	細川	0.0000	0.0000	0.0000	0.0000
53	内閣	0.0087	0.0026	0.0112	0.0043
54	改造	★0.0194	0.0045	0.0162	0.0029
55	騒ぎ	0.0052	0.0005	★0.0156	0.0209
56	外相	0.0028	0.0032	★0.0263	0.0029
57	交代	0.0090	0.0042	★0.0238	0.0108
58	説	0.0105	0.0163	★0.0533	0.0050
59	訪問	★0.0202	0.0055	0.0200	0.0155
60	日程	0.0134	0.0066	0.0142	★0.0157
61	変調	0.0023	0.0008	0.0136	0.0047
62	われわれ	0.0237	0.0071	★0.0256	0.0067
63	側	0.0080	0.0085	★0.0204	0.0033
64	つぶやき	0.0000	0.0000	0.0000	0.0000