雑多なテキスト集合からのストーリー生成

鎌田 健一　　黒橋 禎夫　　西田 豊明
東京大学工学部　　東京大学大学院情報理工学系研究科

1 はじめに

インターネットの普及などにより膨大な量のテキストが電子化されるようになり，家庭からでもそのような情報が利用できるようになってきた．それに伴い，学術文書やビジネス文書だけでなく，日常生活や地域コミュニティーに関連するような，系統立てて書かれていない雑多なテキストを利用する機会も急速に増えてきている．

日常生活や地域の話題に関する場合，ユーザーの要求は明確になっているものばかりではなく，漠然とした興味でしかない場合も多い．そのような場合には，シャープな検索結果を返すというよりは，ある話題についてストーリーを組み立てて提示したり，テキスト集合の中から面白い話題を発見するような処理が必要となる．

本論文ではユーザーから話題が与えられたときに，その話題に関するテキストをグループ化して提示する手法について述べる．また，ユーザーの漠然とした興味を満たすため，テキスト集合中にどのような話題が含まれているかを検出する方法についても述べる．今回提案する手法の全体の流れは図1のようになる．

2 特定の話題に関するテキストの抽出

話題語に関連するテキストを取り出すために，語の共起度を利用する．語の共起度はそれらが出現するテキストの重なり具合を数値化したもので，それらの語同士の関連の深さを示す．よって，共起の強い語を多く含むテキストは元の語と関連の深いテキストであるということができる．

そこで，話題語に対して共起している語を求め，共起度を計算する．この共起度を元に語に対して得点を与え，それらの語を含むテキストに対して得点を与える．この結果大きな得点を得たテキストを取り出すと，話題語と関連の深いテキストが取り出せることになる．

2.1 語の共起度

語 w_i と w_j の共起度 $co(w_i, w_j)$ は次のように定義する．

$$co(w_i, w_j) = \ln \mathrm{DF}(w_i, w_j) \frac{2\mathrm{DF}(w_i, w_j)}{\mathrm{DF}(w_i) + \mathrm{DF}(w_j)} \cdot$$

$\mathrm{DF}(w_i)$ はテキスト集合全体 D での語 w_i の文書頻度，$\mathrm{DF}(w_i, w_j)$ は D において語 w_i と w_j が共に出現するテキスト数である．

$\ln \mathrm{DF}(w_i, w_j)$ は出現頻度による重みであり，出現頻度の低い語同士が偶然同時に出現し，高い共起度を持ってしまうことを防ぐ目的を持っている [1]．

w_i と w_i，すなわちその語自身同士については，co を次のように定義する．

$$co(w_i, w_i) = \ln \mathrm{DF}(w_i)$$

テキストへの得点や話題語への得点を与える語として，一般名詞と固有名詞を用いる．一般名詞はJUMANによって解析された形態素のうち，形式名詞などを除く名詞・カタカナ語・アルファベット語を用いる．固有名詞が他の固有名詞に含まれる場合どちらも固有名詞とみなし，一般名詞が固有名詞に含まれる場合，固有名詞の最終形態素のみ一般名詞とみなす．例えば，「巣鴨大鳥神社」の中には，「巣鴨大鳥神社」「巣鴨」「大鳥神社」という固有名詞と，「神社」という一般名詞が含まれるとする．

2.2 テキストの抽出

上述の共起度を元に，話題となる語 w_i に対してテキスト集合全体 D から関連するテキストを抽出する．

まず，語 w_i との共起度 co が上位10位以内の語には，得点として co を与え，それ以外の語には得点を与えない．次に，各テキストについて得点を持った語を含めばその得点を与える．この得点はその語を含むか含まないかのみで決まり，2回以上含まれていても重複して得点が与えられることは無い．含まれる語によって与えられた得点がテキストの得点となり，得点の上位のテキストが w_i に関連するテキストとして抽出される．

3 テキスト集合全体からの話題の抽出

ここでは，テキスト集合 D のみが与えられたときに，D 中に含まれる話題を表す語を抽出する手法について述べる．

本手法では単に頻度を見るのでなく，「その語と関連の深いテキストが D 中に多く含まれ，それらのテキスト同士の関連も高いような語」を話題として取り出

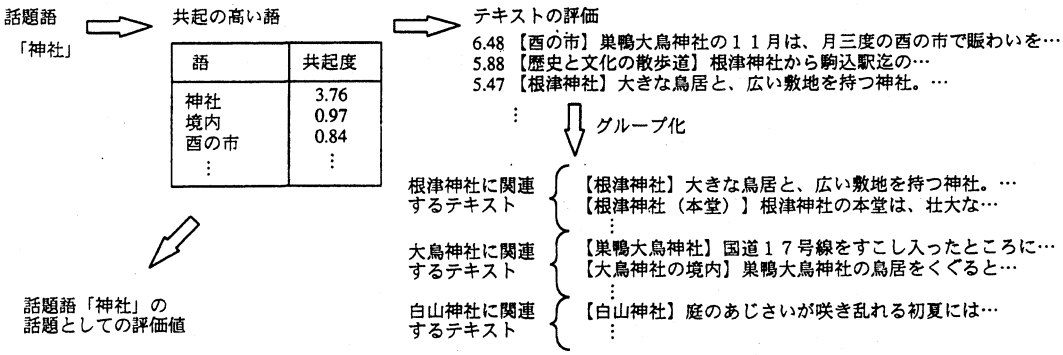

図 1: 提案する手法の概要

すことを考える．これを評価する尺度として，上述した共起度 co を利用する．すなわち，語 w_i に対して共起度 co の高い語上位 10 語の共起度の和を，語 w_i の面白さの評価値とする．

ただし，w_i に対して「樋口一葉」「樋口」「一葉」のように部分一致する固有名詞が共起している場合，最高得点のもの 1 つだけに得点を与える．

4 テキストの並びの生成

2 節の方法で，話題から関連するテキストを取り出し，次にテキスト中に含まれる語によってテキストをグループ化してユーザーに示すことで，ユーザーが把握しやすいストーリーを生成する．

4.1 語の分類

ユーザーが把握しやすいストーリーは，内容的なまとまりが並列的にあるいは順序を持って並んでいるものである．そこで，グループ化に用いる語にはなるべく同じカテゴリーに属する語を選ぶ必要がある．例えば，「神社」に関するストーリーを作る場合には，「根津神社」「大鳥神社」「白山神社」それぞれに関するテキストをまとめる．

一般名詞の分類には，NTT によって作成されたシソーラス [2] を用いた．地域の話題に関するストーリーを生成するという目的のため，シソーラスによる分類をそのまま用いるのではなく，図 2 のようなカテゴリーに当てはめて利用した．

今回研究対象としている，地域の話題に関するテキスト集合のような場合，固有名詞による分類が有効な場合が多い．そこで，神社や公園などの具体的な施設名などを表す固有名詞を検出する．固有名詞は，図 3 のようなカテゴリーに分類した．

固有名詞検出処理では KNP を利用した．KNP によって解析された文節から名詞相当の形態素列を取り

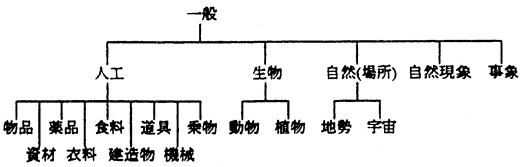

図 2: 一般名詞の分類

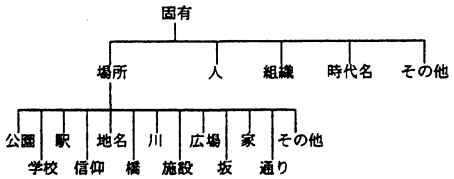

図 3: 固有名詞の分類

出し，その末尾が「界隈」「行き」などの，固有名詞自身に含まれない形態素であれば削除する．その後，形態素列に辞書に載っている固有名詞やカタカナ語が含まれ，形態素列の末尾が「通り」「大学」のように複合語として固有名詞化しやすい語であれば全体を一つの固有名詞とみなす．また，この末尾の形態素を利用して図 3 の分類に当てはめた．

4.2 グループ化のための語の組の評価

抽出されたテキストに含まれる固有名詞・一般名詞を用いて，テキストをグループ化する．このとき，同じカテゴリーに属する語の組は様々な可能性がある．これらの組の中から，各語に属するテキストの重なりがあまり多くなく，評価値の高いテキストを多く取り出せるような語の組を選ぶ．

そのため，図 4 の制約条件を掛けた上で，次の評価値を最大にするような語の組を選択する．

$$\text{出力の評価値} = f(\text{出力テキスト数}) \times \text{出力テキストの評価値の和} \quad (1)$$

$$f(x) = 1/x^{0.7}$$

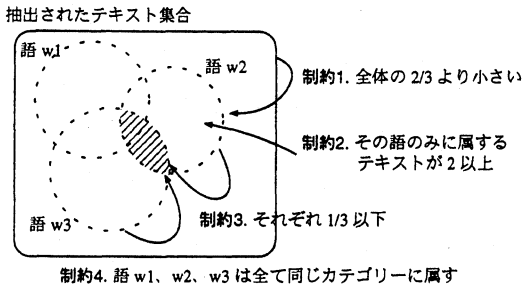

図 4: グループ化における制約条件

一般的に，抽出されたテキストの評価値は図 5(a) のようになる．話を簡単にするために，制約条件が無かったとして評価値の大きなテキストから順に取り出すと，テキストの評価値の累計は (b) のようになる．これに (c) の係数 $f(x)$ をかけると (d) となる．これが最終的な評価値である．これを最大化するようにテキストを取り出すことで，評価値の大きなテキストのみを取り出すことができる．

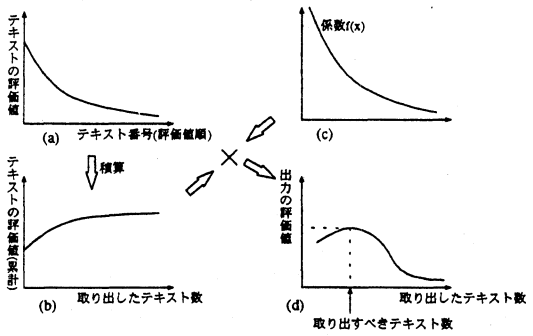

図 5: 各テキストの評価値と出力の評価値の関係

最終出力のグループ同士の並び順は，各グループに属するテキストに式 (1) を適用して決め，グループ内のテキストの順は，グループ化に用いた語がタイトルや第 1 文目に含まれるかどうかで決める．

5 提案する手法の実験結果と考察

本研究では文京区の地域に関するテキスト集合約 560 件 [3] を対象に実験を行った．ストーリー生成の具体例を図 6 に示す．

5.1 話題の検出の評価

今回提案した手法によって検出された話題と，比較対象として，文書頻度が上位の語を表 1 に示す．

表 1: 話題の検出結果

話題	評価値	高頻度語	頻度
東大	8.05	坂	80
坂	7.64	文京	69
本郷	7.63	本郷	56
江戸	7.14	文京区	49
駅	7.09	昔	46
湯島	7.09	東大	45
丁目	6.96	神社	43
小石川	6.91	江戸	39
神社	6.90	東京	36
案内	6.81	白山	35

この結果を評価するために，各語が話題として適切かどうかを調べた．それぞれの語について関連するテキストを上位 30 テキスト抜きだし，そのテキストを読んだ上でその語が文京区の話題として適切かどうかを判断した．話題として適切であったものを○，不適切だったものを×，その語に関連するテキストはまとまっているがその語が話題を表さないものを△として集計したものが表 2 である．この結果から，提案した手法によって語の出現頻度では得られない話題を検出できていることが分かる．例として，頻度 10 の「樋口一葉」，頻度 8 の「綱吉」などが検出できている．

表 2: 話題検出の評価

	今回の手法		頻度によるもの	
	○の数	○と△の数	○の数	○と△の数
上位 10 語	7	8	6	7
上位 20 語	14	16	10	11
上位 30 語	19	23	15	18

5.2 テキストの並びの生成の評価

グループ化およびテキストの出力順の制御によって把握しやすいストーリーが生成されたかどうかを評価した．評価の対象として，提案した手法によって並べられたテキスト，および，並べられていないテキストとしてテキストにつけられた評価値順に上位 15 テキストを取り出したものを用い，両者を比較した．話題となる語には，今回の話題検出手法によって検出された語から，5.1 節の評価で○または△のものを取り出し，固有名詞・一般名詞ごとにそれぞれ上位 10 語を用いた．

両者を比較して，提案した手法によって並びを生成した結果把握しやすくなったものには○，把握しにくくなったものには×，どちらとも言えないものには？をつけて，まとめたものが表 3 である．65 %の話題について把握しやすさの改善が見られた．

「東大」や「駅」のように関連する話題が複数あってグループ化に用いる語の自由度が高い話題では，よい結果が得られている．よいストーリーが得られなかった話題についてはいくつかの原因が考えられる．まず，「江戸」の場合は，関連する話題が複数あるもの

「並木」を話題としたときの出力

【桜】

『桜並木』播磨坂と呼ばれるこの坂は、上から下まで植えられた桜の木で花の季節にはピンクのじゅうたんのようになり、見事です。

『桜並木』環状三号線の中には桜並木があります。１２７本の桜たちは、春になると一斉に開花し、桜祭りの頃には桜の花弁の屋根が出来ます。

『神田川桜並木』自転車歩行者専用道路で、左岸１５４m、右岸３２m、桜２１本、つつじ６３０本、照明灯１１基だそうです。

『播磨坂』春には桜でいっぱいになる播磨坂は、５００メートルも続く有名な桜並木です。

『江戸川公園を臨む』胸突坂の下にある駒塚橋より江戸川公園を臨む。春は川沿いの桜並木が美しい。後ろにフォーシーズンズホテルが見える。

【銀杏】

『銀杏並木』ここでも思い出すのは道の左側の全共闘の拠点の物品館、右側文学部での林健太郎学部長の吊し上げ、加藤一郎総長代理代行の醒めたインタビュー、機動隊にぶん投げられて壊された眼鏡のツル、催涙ガス、キドカラーと揶揄した機動隊制服の青色。

『東大正門前』ここから入ると東大のシンボル、安田講堂へまっすぐ続く銀杏並木の道となります。残念ながら私の二人の息子どもは、この門をくぐることはなく、私立の大学へ行ってしまいました。

「綱吉」を話題としたときの出力

【小石川】

『小石川植物園』かつては後の将軍徳川綱吉の御殿であり（白山御殿），幕府の御薬園を経て東大付属の植物園となる。「赤ひげ」で有名な小石川養生所の舞台であり，青木昆陽の進言で，飢饉の備えとしてサツマイモの栽培，品種改良に努めたのもこの地であった。

『御殿坂の看板』小石川植物園の高い塀沿いに、この看板がありました。ここは徳川５代将軍綱吉の館林城主時代の別邸の地だったそうで、それにちなんで白山御殿と呼ばれたとか。

『御殿坂』小石川植物園の脇を白山の方向に向かって上る坂。かなりきつい勾配。五代将軍綱吉の白山御殿があったためこの名がつき，それ以前は大阪，頂上で富士の山が見えたので富士見坂とも言われた。

『植物園』小石川御殿ともいわれたこの地は将軍綱吉の時代に幕府の薬園となった。明治になり東京大学付属植物園となる。

『白山神社』西暦９４８年，加賀の白山神社を勧請したという，古社。小石川御殿（白山御殿）造営を機に現在の地に移転となった縁で，徳川綱吉とその母君から厚い信仰を受けた。民衆からは，歯痛に効く流行神として迎えられた。紫陽花祭りで有名である。三田線白山駅降りてすぐ。

【湯島】

『湯島聖堂』湯島聖堂内の孔子廟である大成殿。儒教に傾倒した徳川五代将軍綱吉が創建し、自らも「論語」の講釈を行っていた。

『湯島聖堂』１６９０年（元禄３年）五代将軍徳川綱吉によって建てられたもの。いわゆるここは江戸幕府の教育機関。

図 6: 出力の具体例

表 3: ストーリー評価結果

固有名詞		一般名詞	
話題	評価	話題	評価
東大	○	坂	?
本郷	○	駅	○
江戸	?	神社	○
湯島	○	線	?
小石川	○	植物園	○
綱吉	○	昔	○
菊坂	?	館	○
後楽園	○	学生	?
樋口	?	観察	○
白山	?	小学校	○

の，「坂」「神社」「幕府」のようにカテゴリーが違うためにグループ化できず，「江戸時代」と「明治時代」でグループ化されてしまい，江戸に関する話題が散乱してしまった．また，「菊坂」の場合は既に話題が密であったためうまくグループ化できなかった．

5.3 グループ化の種類

グループ化は次の２つに大別できる．

- 話題語の下位の語による分類
- 話題語とは別の軸の概念による分類

一般名詞を話題として与えた場合両者の分類が出現し，また，一般名詞による分類，固有名詞による分類ともに存在した．

しかし，固有名詞を話題として与えた場合，下位の固有名詞による分類がほとんどで，別の軸の固有名詞による分類は人名に対してゆかりの地名で分類されるものがあるだけであった．一般名詞による分類は出現しなかった．これは，一般名詞の場合内容的に並列な名詞が必ずしも同じカテゴリーに属さないことと，固有名詞によって抽出されたテキストは既に内容がかなり限定されていて，内部でのグループ化の自由度が低いためであると考えられる．

6 まとめ

本論文では雑多なテキスト集合からユーザーの把握しやすいストーリーを生成するため，テキストを内容によってグループ化する手法を示した．また，対象テキスト集合のみが利用可能である場合に，その集合に含まれる話題を自動的に検出する手法を示した．このような雑多なテキスト集合を有効に利用するための技術は，インターネットが個人にまで普及し，日常のコミュニケーションの道具となるにつれて，今後ますます重要になっていくと考えている．

参考文献

[1] 北村美穂子, 松本裕治: 対訳コーパスを利用した対訳表現の自動抽出, 情報処理学会論文誌 Vol. 38 Num. 4 pp. 727–736, 1997.

[2] NTT コミュニケーション科学研究所: 日本語語彙大系, 岩波書店, 1997.

[3] 西田・黒橋研: POC コンテンツ・東京都文京区・平成 13 年度版, 東京大学大学院情報理工学系研究科, 2001.