社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－NL－164（26）
2004／11／6

サブストーリーモデルに基づく文章の流れの抽出

砂山　渡†　橘　啓八郎†

† 広島市立大学情報科学部　〒731-3194 広島市安佐南区大塚東 3-4-1
E-mail: †sunayama@sys.im.hiroshima-cu.ac.jp

あらまし　我々はさまざまな生活や活動の局面において，文章を読み書きする機会がある．理解しやすい文章は前後の段落間の関係が明確であり，全体を通して一貫した話の流れが存在する．本研究においては，文章のテーマとその流れ，および文章のサブテーマとその流れをモデル化したサブストーリーモデルについて述べ，サブストーリーを表す単語を評価する方法の提案を行なう．また，既存の文章のセグメントの並びを，提案手法により再生成する実験を行ない，提案手法の有効性について考察する．
キーワード　ストーリー抽出，文章の流れの抽出，サブストーリー

Document Stream Extraction using Sub-Story Model

Wataru SUNAYAMA† and Keihachiro TACHIBANA†

† Faculty of Infomation Sciences, Hiroshima City University　Ozuka-Higashi 3-4-1, Asa-Minami-ku, Hiroshima, 731-3194 Japan
E-mail: †sunayama@sys.im.hiroshima-cu.ac.jp

Abstract　We have many opportunities to read or write documents. Documents that are easy to read and comprehend have flows of their stories, since relationships among two or more succeeding paragraphs are clear and certainly connected. In this paper, we describe about a sub-story model that consists of sub themes and their flows, before we propose a method to evaluate words representing sub-stories. We held experiments to revive the flows of existing stories, then discuss about the proposing method.
Key words　story extraction, document stream extraction, sub-story

1. はじめに

我々はさまざまな生活や活動の局面において，文章を読み書きする機会がある．理解しやすい文章は前後の関係が明確であり，全体を通して一貫した話の流れが存在する．この文章の流れを抽出することができれば，文章に触れるさまざまな場面において，有効な支援が行なえるようになる．

例えば，長い文章を読む際に，文章全体のテーマに加えて，各セグメントにおけるテーマを表す単語とその変遷を提示することができれば，文章の理解に役立てられる．また，文章を作成する際にも，自分の作成した文章の流れを確認することができれば，文章の記述の改善に役立てることができる．

これらの他に，複数テキストの要約において，要約結果の提示順序を考える際に，要約元のテキスト集合や要約文の集合からストーリーの流れを生成し読みやすい要約の生成を行なうことや，多くのテキストを読み進めながら学習を行なう知識ナビゲーションにおいて，情報提供のプランニングにストーリーの流れを考慮するなどの応用が考えられる．

そこで本研究においては，文章の主題（テーマ）とその流れ，および文章のサブテーマとその流れをモデル化したサブストーリーモデルについて述べ，サブストーリーを表す単語を評価する方法の提案を行なう．

2. 研究背景

文章やセグメント等のテキスト間の関係を与える指標として最もよく用いられているものの1つに，テキストを，出現する単語のベクトルとして表し（ベクトル空間モデル [1]），ベクトル間の角度によりテキスト間の類似度を表す cos（コサイン）類似度がある．

この cos 類似度には次の2つの特徴がある．

1. 2つのテキスト間の静的な関係を表す
2. 2つのテキスト間の関係しか測れない

1. の点は，cos 類似度は2つのテキスト間の関係をその各々が含む単語に基づいて定め，関係を測る2つ以外に存在するテキストその他の情報を，関係に影響させないことをいう．動的な関係を測る従来研究としては，ある観点に基づいてベクトル

内の単語の重みを変化させるもの[2]などがある．

また，2.の点に関して，ベクトル間の角度が2つのベクトルから測られる数値であるため，cos類似度には2つのテキスト間の関係しか数値に現れてこない．

すなわち連続する2つのテキスト間の関係を静的に測るcos類似度のみによって，連続する3つ以上のテキストにかかわるストーリーを評価することは難しいと考えられる．

1つの文章内のセグメントを対象とする研究にも，単語の共起関係やシソーラスを用いた語彙的連鎖[3]の応用したテキストセグメンテーションや，テキスト中での話のつながりをText-tiling法で検出してトピック抽出を行なう研究[4]がある．同じ単語の共起のみを扱う本研究は，シソーラスを使わない語彙的連鎖とも考えられるが，本研究では，各1つ1つの単語が部分的なストーリーの主題になれるという立場を取り，数多くのサブストーリーを扱う点が異なる．また，Text-tiling法においては隣接する2つのブロック間の関連度を，ベクトル空間モデルによるcos類似度で測っており，本研究は隣接する2つ以上のブロックを評価する点が異なる．

文章の流れを保持する要約の研究[5]もある．この研究においては，文章の本筋を隣接する文や段落間の関係から抽出しているが，元の文章内における単語の出現の順序関係を手がかりに導入部と結論部の計算を行なっており，前後の順序関係を含めたストーリーの流れを導出するには至っていない．

テキスト集合における各テキストの因果関係を基に，あるテキストの結論部分と次のテキストの原因部分とを結び付ける方法によって，テキスト間でのストーリーを生成する研究[6]もある．しかし，個々のテキストが原因と結果により完結する内容であるとは限らないため，ストーリーの一部としてテキストを位置づけるには至っていない．

そこで本研究においては，テキスト集合の各テキストに対して，ストーリーの流れの中における各テキストの位置づけを与え，順序付けを行なうためのモデルとそのモデルに基づくストーリー生成手法を提案する．

3. サブストーリーモデル

本章で，1つの文章内に現われるメインストーリーとサブストーリーの定義を述べた上で，文章の流れを表すためのサブストーリーモデルを定義し，このモデルに関する妥当性を示す予備実験について述べる．

3.1 サブストーリーモデルの定義

文章には全体を通じて筆者が述べたい主な話題（メイントピック）と，主題に関連して文章の一部で述べられるサブトピックとがある．そこで，各トピックを代表するキーワードを以下のように定める．

1. トピックキーワード：文章のメイントピックを表すキーワード
2. サブトピックキーワード：文章のサブトピックを表すキーワード

また，トピックキーワードに関する話をメインストーリー，サブトピックキーワードに関する話をサブストーリーと呼ぶ．

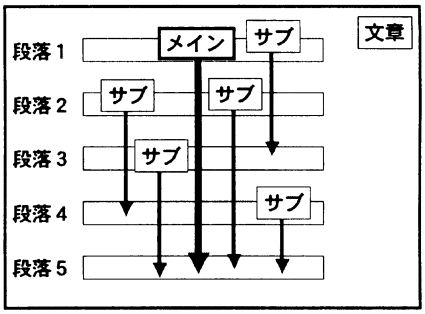

図1 サブストーリーモデル

表1 セグメント数の分布

SEGMENT	論文数	平均単語種類数
3	2	77
4	15	104
5	45	115
6	30	119
7	3	181
8	4	88
9	1	117
合計／平均	100	115

これらで定義されるサブストーリーモデルを図1に示す．ただし，図中の「メイン」と「サブ」はそれぞれメイントピックとサブトピックを表し，矢印はストーリーを表す．すなわち文章には，主題に関わる大きな話の流れが存在すると共に，文章の各部分での話題に伴う局所的な話の流れが存在する．

本モデルにおいては，数文から数十文の固まりを話の流れの一つの単位として想定している．これは，サブストーリー同士の関連を明確にする目的において，個々の単位がある程度まとまった量の単語を含む必要があることと，大きすぎる単位では関連を詳細に把握できないことによる．

3.2 サブストーリーモデル予備実験

3.1で定義したサブストーリーモデルがどれだけ客観的なモデルとなっているかを，確かめる予備実験を行なった．すなわち，局所的に存在するサブストーリーのテーマとなり得る単語が，文章中に少なからず存在することを実験により確かめる．実験に用いたテキストは，第18回人工知能学会全国大会の論文集から抽出した4ページの論文100件であり，第1章から結論の章までの各章をひとつのセグメントとみなし，図や数式などの記述は予め取り除いてある．抽出した単語は，名詞，動詞，形容詞のうち「ある，いう，みる」などの形式詞を除いたものとした．表1に，100件の論文のセグメント数(SEGMENT)と単語数に関するデータを示す．

この用意したテキストデータに対して，同じ単語がどれだけ連続するセグメントに出現しやすいかを調べた結果を表2に示す．ただしデータは，論文数が多いセグメント数(SEGMENT)が4から6のもののみを示す．この表より，全ての実験値は理論値を大きく上回っており，単語が連続するセグメントに出現

表 2　セグメント頻度 (SF) が n の単語が，連続する n 個のセグメントに出現する確率

SEGMENT	SF	全単語数	連続単語数	確率	理論値
4	2	981	719	73%	50%
5	2	3025	1728	57%	40%
6	2	1875	917	49%	33%
4	3	368	254	69%	50%
5	3	1230	561	46%	30%
6	3	860	310	36%	20%
5	4	602	286	48%	40%
6	4	468	176	38%	20%
6	5	235	88	37%	33%

表 3　セグメント頻度 2 の単語が出現するセグメント区間幅の累積割合

SEGMENT	4	4	5	5	6	6
区間幅	理論値	実験値	理論値	実験値	理論値	実験値
2	50%	73%	40%	57%	33%	49%
3	83%	95%	70%	85%	60%	72%
4	100%	100%	90%	95%	80%	88%
5			100%	100%	93%	96%
6					100%	100%

表 4　セグメント頻度 3 の単語が出現するセグメント区間幅の累積割合

SEGMENT	4	4	5	5	6	6
区間幅	理論値	実験値	理論値	実験値	理論値	実験値
3	50%	69%	30%	46%	20%	36%
4	100%	100%	70%	77%	50%	66%
5			100%	100%	80%	87%
6					100%	100%

しやすくなっていることが分かる．

また，全ての同じ単語が連続して出現しない場合において，出現するセグメントの範囲を調べた結果を表 3 表 4 に示す．これらの表より，すべての実験値は理論値を上回っており，各単語は全てが連続するセグメントに出現しない場合においても，できるだけ連続するセグメントに出現しやすい傾向にあることが分かる．

以上の予備実験により，単語は部分的に固まって出現する傾向にあることがわかった．すなわち，何らかの意図に基づいて文章中に局所的に単語が出現すると考えられるため，この単語をサブストーリーの主題とするサブストーリーモデルの妥当性が確認された．

4. サブストーリーモデルに基づく文章の流れの抽出アルゴリズム

4.1 トピックキーワードの抽出

本研究においては，トピックキーワードを文章全体を通じて現れるキーワードとし，サブトピックキーワードを，文章の一部分において現れるキーワードと定義する．

そこでまず，繰り返し複数のセグメントに出現する単語を評価することで，(サブ) トピックキーワードの抽出を行なう．文章中の各単語 $word$ のセグメント頻度 $SF(word)$ (単語が出現するセグメント数) を数え，これが 2 以上の単語をサブトピックキーワードとして抽出する．

次に，メインとサブのトピックキーワードを区別するために，各単語の出現する区間の長さ $R(word)$ を式 (1) で与える．ただし End と $Start$ はそれぞれ，キーワードが最後に現れるセグメントと最初に現れるセグメントの番号を表す．

$$R(word) = End(word) - Start(word) + 1 \tag{1}$$

すなわち，最初のセグメントと最後のセグメントの両方に現れる単語の $R(word)$ は，文章のセグメント数を N とするとこれに等しくなる．そこで，しきい値 K を用いて，$R(word)$ が総セグメント数 N のおよそ 80%以上の値を持つ単語をメイントピックキーワード，$R(word)$ が 2 以上 $N \times 0.8$ 未満の単語をサブトピックキーワードとする．

4.2 サブストーリーの評価手法

サブストーリーを評価する手法を，以下で述べる 4 種類の方法で実現した．

4.2.1 出現密度の最大化による方法

各単語 $word$ に関して，ストーリーの流れの評価を式 (2) で行なう．すなわち，単語の出現するセグメントの区間における出現密度とセグメント頻度とをかけ合わせることで，サブストーリーとしてより長く，より連続的に現れる単語を評価する．

$$\begin{aligned} Sub(word) &= \frac{SF(word)}{R(word)} * SF(word) \\ &= \frac{SF(word)^2}{R(word)} \end{aligned} \tag{2}$$

また，文章の流れの評価値を各単語のサブストーリーの評価値の和として式 (3) で定める．

$$Stream(doc) = \sum_{word \in doc} Sub(word) \tag{3}$$

4.2.2 出現区間最小化による方法

出現区間最小化は，各単語 $word$ の出現するセグメントの範囲を評価し，式 (4) を用いてより短い区間により出現するサブトピックキーワードに高い値を与える．その後，文章の流れの評価値を式 (3) を用いて与える．ただし，N は文章 doc のセグメント数である．

$$Sub(word) = N - R(word) \tag{4}$$

4.2.3 cos 類似度を用いた方法 1

まず任意の 2 つのセグメント間の cos による類似度を，セグメント頻度 2 以上の単語を用いて，式 (5) により計算する．ただし，式中の $Common(P_i, P_j)$ はセグメント P_i と P_j に共通して現れる単語の数，$Num(P_i)$ はセグメント P_i に現れる単語の数を表す．また，N は文章 doc のセグメント数である．

$$cos(P_i, P_j) = \frac{Common(P_i, P_j)}{\sqrt{Num(P_i) * Num(P_j)}} \tag{5}$$

文章の流れの評価値を，距離 d だけ離れたセグメントの類似度の和を用いて，式 (6) で与える．ただしセグメント間の距離

表 5　文章再構成実験結果の順位による累積頻度（先頭のセグメント固定）

	DEN	ST	COS	理論値
1位	28	25	22	6
2位	43	45	36	11
3位	48	46	43	16
4位	56	51	51	20
5位	61	57	59	25

は，隣接するセグメント間の距離を1として数える．また，k は流れの評価を行なうセグメント間の距離の上限 (N) に対する割合 $(0 < k \leqq 1)$，r は距離の短いセグメント間の類似度を優先するための重み $(0 < r \leqq 1)$ である．

$$Stream(doc) = \sum_{d=1}^{k*N} \sum_{i=1}^{N-d} cos(P_i, P_{i+d}) * r^{d-1} \qquad (6)$$

5. サブストーリーモデルに基づく文章再構成実験

既存の文章をセグメントごとに切り分け，特定のセグメントの位置を入力として与え，残りのセグメントを全体として最も流れが良くなるように自動的に並べて文章を再構成する実験を行なった．これにはまず，計算機により自動的にセグメントの並びの全パターン(注1)．を生成した上で，値の最も高い並びをシステムの出力とした．

実験は，以下の評価関数を用意して比較することで評価を行なった．

1. 出現密度の最大化による手法 (DEN)
2. 出現区間最小化による手法 (ST)
3. cos 類似度を用いた手法 (COS) (注2)

実験評価は，正解をもとの文章におけるセグメントの並びとし，各評価手法による評価値による順位付けにより，正解が第何位に出力されるかを比較することで行なった．実験に用いたテキストは，第18回人工知能学会全国大会の論文集から任意に抽出した100件であり，第1章から結論の章までの各章をひとつのセグメントとみなし，図や数式などの記述は予め取り除いてある．

5.1 実験1：緒論からのストーリー生成

まず，先頭のセグメントのみを入力として与えたときに，残りのセグメントを自動的に並べ変えて，もとの論文と同じに再生できるかを確かめる実験を行なった．すなわち，ストーリーの緒論を入力として与えた時に，最も自然な流れとしてもとのストーリーが生成できるかを確認する．このときの実験結果を表5に示す．

この表によると，いずれもセグメントをランダムに並べた際の理論値を大きく上回っており，単語の連続的な出現を仮定するサブストーリーモデルの効果が現れている．手法間の比較においては，DEN,ST,COS の順に良い結果となっている．これは，単語出現の連続性をより強く評価する順であり，COS は同じ単語が3セグメント以上に連続で出現する場合を直接的に測ることができないため，このような結果になったと考えられる．

(注1)：セグメント数が N の場合は $N!$ 通りの並びのパターンが生成されるが，あらかじめ入力として位置を指定したセグメントの数 P に応じて，$(N-P)!$ 通りに減少する．

(注2)：式 (6) のパラメータは，実験的に k=0.8,r=0.5 として定めた．

表 6　文章再構成実験結果の順位による累積頻度（最後のセグメント固定）

	DEN	ST	COS	理論値
1位	14	13	9	6
2位	27	28	18	11
3位	33	33	25	16
4位	37	34	31	20
5位	41	38	35	25

表 7　文章再構成実験結果の順位による累積頻度（先頭と最後のセグメント固定）

	DEN	ST	COS	理論値
1位	34	33	42	18
2位	60	60	64	35
3位	66	65	72	43
4位	71	71	78	52
5位	75	76	80	61

5.2 実験2：結論からのストーリー生成

次に，最後のセグメントを入力として与えたときに，その結論を導くストーリーを再生する実験を行なった．実験結果を表6に示す．

先の実験と同様に，いずれのシステムも理論値を上回る結果を出しているものの，その出力の精度が大きく下がっている．これは，論文の結論部分があまり長くないことにより，結論のセグメントと他のさまざまなセグメントが連結しやすい状況にあったためと考えられる．また実験1と同様の理由で，DEN,ST,COS の順に良い結果となった．

5.3 実験3：緒論と結論からのストーリー生成

次に，先頭と最後の両セグメントを入力として与え，間のセグメントを補完して再生する実験を行なった．このときの実験結果を表7に示す．

いずれの指標においても理論値を大きく上回っている．手法間の比較では，実験1，2と異なり，COS が ST と DEN よりも良い結果となった．これは，先頭や最後のセグメントはどのセグメントとも少なからず共通点をもつという，曖昧性の高いセグメントが固定されたため，単語出現の連続性よりもむしろ，単純に前後のセグメントのつながりが，論理展開の再生に重要であったためと考えられる．

5.4 本稿テキストを用いたストーリー再生実験

本節では，本稿を用いて行なったストーリー再構成実験について述べる．本稿の本節を除くテキストを各章ごとにセグメントごとに切り分け，先頭および最後に来るセグメントを入力として与え，残りのセグメントを全体として最も流れが良くなるように自動的に並べて文章を再構成する実験を行なった．テキストは7セグメントからなり，273種類の単語を含む．各単語のセグメント頻度の分布を表8に示す．

表 8　本稿の単語のセグメント頻度の分布

セグメント頻度	単語数	累計割合（%）
7	5	1.8
6	4	3.3
5	4	4.8
4	8	7.7
3	27	17.6
2	60	39.6
1	165	100.0

表 9　本稿テキスト再生実験における正解の並びの出力順位

固定ポイント	DEN	ST	COS
先頭と最後	8/120	14/120	19/120
先頭	13/720	23/720	90/720
最後	29/720	69/720	153/720

表 10　DEN の出力結果（先頭と最後固定：上位 10 件）

順位	セグメントの並び	評価値
1	1 2 4 6 5 3 7	249.29
2	1 2 4 3 5 6 7	246.58
3	1 2 6 4 5 3 7	245.14
4	1 2 6 3 5 4 7	245.04
5	1 2 3 6 5 4 7	244.61
6	1 2 6 5 3 4 7	244.43
7	1 2 3 5 6 4 7	242.49
8	1 2 3 4 5 6 7	241.99
9	1 3 5 6 4 2 7	240.67
10	1 6 5 3 4 2 7	240.41

セグメント頻度 7 の単語は「文章，流れ，セグメント，単語，ストーリー」，頻度 6 の単語は「サブ，抽出，出現，連続」となっており，これらの単語は 7 セグメントの 8 割以上に出現する単語として，メインストーリーのテーマを表す単語とする．また，セグメント頻度 5 の単語には，「モデル，テキスト，全体，評価」，頻度 4 の単語には「確認，前後，存在，結果，文，類似，部分，実験」という単語がある．これらの単語を初めとし，頻度 5 以下の単語はサブストーリーのテーマを表す単語として，それぞれに関するストーリーがあると考える．

再構成を行なうシステムは，先の実験と同様に DEN,ST,COS の 3 種類である．本稿のセグメントの並びを正解とした時に，各システムにおける正解の並びの評価値の順位を表 9 に示す．この結果によると，DEN,ST,COS の順に良い結果が得られている．セグメントの並べ方の全組み合わせの数である 120 と 720 に比較すると，比較的よい結果を出している．これは 3.2 節の予備実験の結果にもあるように，本文中の単語が連続するセグメントに出現しやすい傾向にあったためである．

しかしどのシステムも，正解を上位 5 位以内に出力するには至らなかった．表 10 と表 11 に，先頭と最後のセグメントを固定した実験と，先頭のセグメントのみを固定した実験において，最も結果の良かったシステム DEN の上位の出力結果を示す．ただし，表中の「セグメントの並び」の項の数字は，本稿の各章に対応するセグメント番号を並べたものである．すなわち，正解の並びは「1234567」となる．

表 11　DEN の出力結果（先頭固定：上位 13 件）

順位	セグメントの並び	評価値
1	1 2 4 6 5 3 7	249.29
2	1 2 4 3 5 6 7	246.58
3	1 7 3 5 6 4 2	245.72
4	1 2 6 4 5 3 7	245.14
5	1 2 6 3 5 4 7	245.04
6	1 7 6 5 3 4 2	244.84
7	1 2 3 6 5 4 7	244.61
8	1 2 6 5 3 4 7	244.43
9	1 2 7 6 5 3 4	244.30
10	1 2 7 3 5 6 4	242.94
11	1 2 3 5 6 4 7	242.49
12	1 7 2 6 5 3 4	242.29
13	1 2 3 4 5 6 7	241.99

表 12　本稿テキスト再生実験における正解の並びの出力順位（改良後）

固定ポイント	DEN	ST	COS
先頭と最後	4/120	13/120	14/120
先頭	5/720	16/720	68/720
最後	18/720	60/720	114/720

先頭と最後のセグメントを固定した，正解が 8 位の結果（表 11）に比べて，先頭のセグメントのみを固定した実験結果（表 10）においては，3,6,9,10,12 位に第 7 セグメントが 2,3 番目にある並びが，正解を上回る評価値を得ている．これは，論文など最初と最後に同じことが述べられる型の文章において，最初と最後のセグメントがつながりやすくなったためである．したがって，取り扱う文章の型に応じて固定するセグメントを定めることで，結果が大きく改善される場合があると考えられる．

表 10 や表 11 における典型的な間違いのパターンを見ると，第 3 セグメントと第 5 セグメントが結びつく傾向にあることがわかる．本文を見ると，3 章ではサブストーリーモデルに関する予備実験について述べ，5 章ではストーリーの再構成に関する実験を述べており，両章が 4 章を挟んで実験に関して述べていることが原因であると考えられる．

また今回，複合語については考慮していなかったため，両者に出現する「人工知能学会全国大会」という本来 1 つの単語が 5 つに切り分けられ，評価値が 5 単語分加算されている状態であることがわかった．

そこで仮の措置として，「第 18 回人工知能学会全国大会」を一つの単語として置換して再実験したところ，システムの出力が表 12 のように改善された．そこで，今後このような複合語に関する処理は加えていく予定である．

6. 考　　察

単語出現の連続性に関して，明確な論理展開の中に位置付けられるセグメントは，前後のセグメントとのつながりが強いと考えられるため，cos 類似度による指標で流れを再生できるが，逆にさまざまなセグメントと関連して位置づけられるセグメントは，直前や直後のつながりだけでストーリーを再生するには情報が不足しており，より長い区間における出現単語の連続性

を評価することが必要と考えられる．したがって，今後各セグメントに関して，前後の論理展開が強い部分であるか否か，各セグメントの文章中での位置づけを基に，ストーリー再生の手法を切り替えを行なっていく．

また，システムの出力は理論値よりも大きいものであったが，正解が上位に出力されない場合も多く見受けられた．これには，システムと筆者それぞれの側の原因として以下の点が挙げられる．

1. 評価されるべきでない単語が評価されている．

2. 論文自体の文章の流れが良くない．もしくは，筆者が考えるよりも良い流れが存在する．

3. 同じ単語の繰り返し以外による，より強い文章の流れが存在する．

1. はシステム側の問題で，正しく複合語が抽出されていなかったり，文章の流れを推定するにふさわしくない，ノイズとなる単語が含まれていて正しい評価値が得られない場合がある．この点に関しては，さまざまなテキストに対する実験結果の各単語の評価値を精査することで改善を図っていく．具体的には，各単語 $word$ の重み $W(word)$ を設定した上で，文章の流れの評価値を式 (7) のように定める．

$$Stream(doc) = \sum_{word \in doc} W(word) * Sub(word) \qquad (7)$$

今回の実験に際しても，単語 $word$ の出現区間内における平均頻度を重みとする，式 (8) の値を試験的に用いた．これは，頻度が高い単語ほど重要である可能性が高いと考えたためである．

$$W(word) = \frac{FREQ(word)}{R(word)} \qquad (8)$$

しかしながら，5. で行なった実験結果とほぼ同程度の結果しか得られなかった．各単語の評価値をもとに考察を行なったところ，確かに頻度の高い単語は，頻度の低い単語よりも連続しやすい傾向にあった．しかし，流れの評価に用いられる単語の大半は，表 8 に表れているように頻度が低い単語であり，多くの頻度の低い単語が全体として文章の流れを作っているため，頻度の低い単語を単純に軽視することはできないことがわかった．

2. はまとまりがない文章であるなど筆者側の問題で，文章の流れが今ひとつの文章が入力された場合である．このような場合には，本システムの出力結果を筆者にフィードバックすることにより，段落の順序の入れ替えや，連続して出現していない単語を積極的に使うように進めるなど，文章の改善を促すことによって，結果の改善が可能な部分もあると考えている．

3. はさまざまな状況が考えられるが，たとえば表記が全く同じ単語のみをシステムが評価の対象としているため，表記の揺れや言い換えなどに対処できない場合，特定の少数の単語によってのみ流れが作られている場合，文章の複雑な構成により，単語が連続的には出現しないなどの場合が考えられる．

これらの場合，表記の揺れや特定の流れを作る単語への対処方法として，現在は単語のセグメント内の共起のみを扱っているため，単語の文内共起の情報を用いて単語間の関係により単語の重みを与えることや，セグメントの並びに関する制約条件を，他の評価指標から算出して組み合わせるなどの改善が考えられる．

7. 結　　論

文章において，同じ単語が連続するセグメントに出現しやすいという性質に基づく，サブストーリーモデルを提案した．また，本モデルを用いて，既存の文章の流れを抽出して文章を再構成する実験を行ない，その効果があることを確認した．今後の課題としては，本モデルを利用した文章の流れの抽出精度の向上，および，検索結果やある領域の Web テキストの集合に流れを作り，読みやすい要約や新しいストーリーの生成などに役立てたいと考えている．

謝辞　本研究の一部は，科学研究費補助金，課題番号 (16700150) の援助を受けて行なわれた．

文　　献

[1] Salton, G, A. Wong, and C. S. Yang: "A Vector Space Model for Automatic Indexing", Communication of the ACM, Vol.18, No.11, pp.613 – 620, (1975).

[2] 笠原要, 松澤和光, 石川勉：国語辞書を利用した日常語の類似性判別，情報処理学会論文誌，Vol.38, No.7, pp.1272 – 1283, (1997).

[3] Morris, J. and Hirst, G.: "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", Computational Linguistics, Vol.17, No.1, pp.21 – 48, (1991).

[4] Hearst, M.A: "Multi-paragraph segmentation of expository text", Proceedings of the 32nd conference on Association for Computational Linguistics, pp.9 – 16, (1994).

[5] 市丸夏樹，飛松宏征，日高達：話題の流れを保持する自動要約，第 160 回情報処理学会自然言語処理研究会資料，pp.43 – 48, (2004).

[6] 大澤幸生，村上尚央，谷内田正彦：内容における因果関係を用いた文書集合からのストーリー抽出，第 33 回人工知能学会 SIG-FAI 研究会資料，pp.43 – 48, (1998)