1B1-01

# テキストの話の流れを視覚化するインタフェース

Text Stream Visulalization Interface

砂山　渡
Wataru SUNAYAMA

広島市立大学大学院　情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

We have many occasions to read electrical texts along with the growth of computers and the internet. An environment to grasp whole contents and flows is required when we comprehend those texts quickly. Automatic summarization methods are also used for this purpose but indicative summaries require our reading whole texts after using them. Since informative summaries also have some quantity, the simple architecture to know the whole texts become useful. In this study, the system labels each word in a text according to the theme, and visualizes labeled words on the interface to know when and how each word is related to the theme.

## 1. はじめに

コンピュータとインターネットの普及に伴い，電子テキストを獲得することは容易になってきたが，人間がそれらのテキストを読んで処理する能力が変わるわけではないため，得られたテキストの取捨選択や，読むべきテキストを素早く閲覧，理解するための環境が望まれるようになってきた．

テキストの内容を素早く知るためには，その要約が用いられることが多いが，指示的要約は興味があるテキストへのポインタとなることが主な役割であり，テキストの内容全体を俯瞰する目的で使用するには不十分と考えられる．また，テキストの内容全体を知るための報知的要約は一般に情報の圧縮率が低く，もとのテキストの少なくとも 30%以上を読む必要が生じる [1] ため，テキスト全体の内容を素早く知るという目的において，より効果的な環境が望まれる．

そこで本研究では，テキスト中の各単語に，テキストの主題との関係を表すラベルを付与し，この主題との関係を明確にする視覚化インタフェースを提案する．すなわち，ラベル付けされた各単語，および各ラベルが与えられる単語の総数に着目し，テキスト全体として，主題に関係する単語がどこでどの程度使われているかを把握でき，主題に関する一貫性の存在の有無を直感的に確認できるインタフェースを提案する．

## 2. 川下りシステム

本章では，話の流れの理解を支援するための川下りシステム（図 1）について説明する．

川下りシステムでは，電子テキストとその主題を表す単語集合（観点語と呼ぶ）を入力として，テキスト中で使われている単語に，テキストの主題との関係を表すラベルを付与した上で，テキストの途中位置までにおけるラベル付けの状況を表すシーンを生成する．それら各シーンの集合をもとに，視覚化インタフェース上で，特定のシーンの表示したり，シーンを連続的に変化させるアニメーションを再生することにより，ユーザがテキストの話の流れを理解することを助ける．

連絡先: 砂山渡，広島市立大学大学院情報科学研究科，731-3194 広島市安佐南区大塚東 3-4-1, sunayama@hiroshima-cu.ac.jp

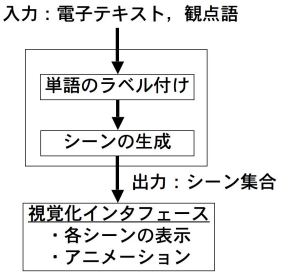


図 1: 川下りシステムの構成

### 2.1 電子テキストと観点語の入力

本稿で扱うテキストは，複数のセグメント（段落などのテキスト内での意味の区切り）から構成されており，各セグメントは単一もしくは複数の文から構成されているとする．また，各セグメントや文の区切りは特定できると仮定する．観点語は，テキストに含まれている単語のいずれかを，ユーザ自身が与えることもできるが，ユーザが観点語を与えない場合においても，自動要約システムのひとつである展望台システム [2] によって自動的に抽出して与えられる．展望台システムは，一文内での単語の共起性をもとに，多くの単語と同時に現れる単語を観点語として抽出する．すなわち，テキスト全体を通じて出現し，最も一貫性がある単語を抽出するため，テキストの話の流れを表す本システムの観点語として適切である．

### 2.2 単語間の距離とテキストの主題との関連

テキスト中の各単語*1 にラベル付けを行なうための準備として，単語間の類似度を表す単語間距離を定義する．まず，セグメント $Seg$ 内における単語 $w_i, w_j$ 間の距離（セグメント内距離）を式 (1) で定義する．

*1 なお以下の本稿では「単語」として，形態素解析器 [3] を用いて抜き出される「名詞」を対象としている．これは，一般的にテキストに与えられるキーワードが名詞であることと，単語の認識と理解のしやすさを重視したことによる．ただし，目的に応じて「動詞」「形容詞」など他の品詞を追加することは可能である．また同様の理由で，本稿の実験時には，一文字の単語および平仮名で始まる単語は除いた．

$$distance(w_i, w_j, Seg) = \min_{s \in Seg} |ap(w_i, s) - ap(w_j, s)| \quad (1)$$

ただし $ap(w, s)$ は，単語 $w$ が文 $s$ 内で出現した箇所が，文の先頭から何単語目であるかを与える関数とする．すなわちセグメント内距離を，セグメント内の各文 $s$ において単語 $w_i$ と単語 $w_j$ が出現した箇所間の距離（$w_i, w_j$ 間の単語数）の最小値[*2]として定める．

単語間の距離を，シソーラスなどによって与えない理由は，同じ単語同士でもテキストが異なれば，単語間の類似性も異なると考えたためである．

各テキストには，そのテキストで述べたい主題が存在する．本稿では，後述する単語のラベル付けの際に，各単語が主題に関係しているか否かを重視する．そこで，本節で定義したセグメント内距離を用いる．すなわち，各セグメントにおいて，テキストの主題を表す観点語の集合 $T$ に含まれるいずれかの単語とセグメント内距離が，しきい値 $dmax$ 以下の単語を主題との直接の関係が理解できる「TOPIC 関連語」，しきい値 $dmax$ より大きい単語を主題との関係が明らかでない「TOPIC 非関連語」とする．また現在，$dmax$ は 10 としている．結果として，観点語と同一文中に現れ，かつ観点語と $dmax$ 単語以内に出現する単語が主題に関係する単語となる．

### 2.3 単語のラベル付け

テキスト中の各単語には，テキストの主題との関係を表すラベルを与える．

以下に，テキスト内の各単語に与えられる 6 つのラベルとその意味を示す．

- TOPIC：テキスト全体の観点となる単語．観点語．人手により与えるか，展望台システムで自動抽出する．意味：テキストの主題を表す単語．
- SEED：観点語の近くに，初めて現れた単語．意味：主題に関する話を広げる際に，話の種として用いられる単語．
- MAIN：観点語の近くに現れて，主題に関する論理を展開する単語．意味：話の本筋に強く関わる単語．主題を構成する単語．
- SUB：観点語の近くに現れた後，観点語から離れて現れた単語．意味：主題に関する話に用いられた後，副主題として展開した話にも関係する単語．話の副主題を構成する単語．
- NEW：観点語から離れた場所に，初めて現れた単語．意味：1) 主題の一部分を深く掘り下げた話 (副主題) を広げる際に用いられる単語．2) 主題と関係のない話に用いられる単語．
- BYWAY：一度も観点語の近くに現れないまま，複数回使用されている単語．意味：1) 副主題にのみ関係して，繰り返し使われている単語．冗長な単語の可能性がある．2) 主題と関係のない話に繰り返し用いられている単語．

ラベル付けは，テキストの前から 1 セグメントごとに以下の手順で行なう．

1. 単語が観点語であれば TOPIC のラベルを与える．

*2 単語 $w_i$ と単語 $w_j$ の片方でも出現しないときは無限大とする．

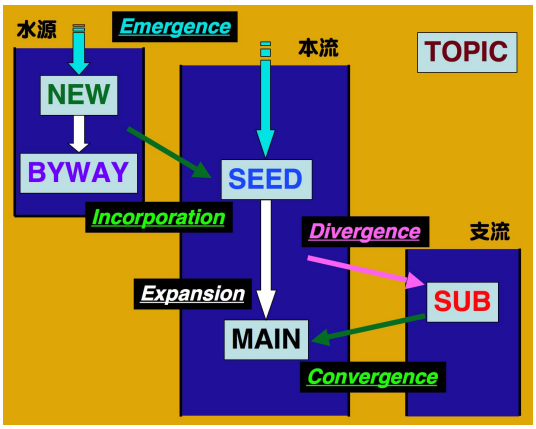


図 2: 同一単語に与えられるラベルの遷移パターン

2. 単語間の距離を表す式 (1) によって，セグメント内の各単語が，観点語と距離 $dmax$ 以内の TOPIC 関連語であるか否かを調べる．
3. テキスト中で初めて出現した単語が TOPIC 関連語であれば SEED，そうでなければ NEW のラベルを与える．
4. テキスト中で 2 回目以上の出現となる単語が TOPIC 関連語で，今までに一度も TOPIC 関連語になったことがない単語には SEED，過去に TOPIC 関連語になったことがある場合には MAIN のラベルを与える．
5. テキスト中で 2 回目以上の出現となる単語が TOPIC 非関連語で，今までに一度も TOPIC 関連語になったことがない単語には BYWAY，過去に TOPIC 関連語になったことがある場合には SUB のラベルを与える．

### 2.4 単語ラベル遷移アーク

同一単語の与えられるラベルが，各セグメント間でどのように遷移するかを図 2 に示す．この遷移関係をアークと呼び，ラベル間の遷移を表す 7 つのアークとその意味を以下に示す．

- Emergence（未出現→ SEED）：主題に沿った新たな単語が現れること．
- Expansion（SEED → MAIN）：主題に沿って一度しか現れていなかった単語を用いて，話を膨らませた．
- Incorporation（NEW or BYWAY → SEED）：観点語から離れて現れていた単語が，初めて観点語の近くに現れること．主題との関係が明確でなかった単語を，初めて観点語を共に用いて話をしたことにより，主題との関係を明らかにした．一見関係のない話や，とりとめのない話の中から，話題のヒントが現れた状況にも相当する．
- Divergence（SEED or MAIN → SUB）：観点語の近くに現れていた単語が，観点語から離れて現れること．観点語の出現頻度を相対的に下げて，主題に関係するある単語について，深く掘り下げた話を始めること．話の発散．
- Convergence（SUB → MAIN）：観点語から離れて現れていた単語が，再び観点語の近くに現れること．一時的に深く掘り下げた話をしていたため観点語の近くに現れなかった単語を，再び観点語と結びつけて話をしたこと


図 3: 川下りシステムインタフェース

により，掘り下げた話の結論と主題との関係を明らかにした．話の収束．

- Out-Emergence（未出現→ NEW）: 主題との関係が不明な新たな単語が現れること．
- Out-Expansion（NEW → BYWAY）: 主題との関係が不明な一度しか現れていなかった単語を用いて，主題との関係が不明な話を膨らませた．

これらのラベル遷移アークは，次章で述べる視覚化インタフェース上で，ユーザがアニメーションを見て，テキストの話の流れ，特に各セグメントの話の流れの中での位置付けを理解するために用いられる．

## 3. 視覚化インタフェース

本章では，前章で定義した，テキスト中の各単語に与えられるラベルと，ラベルの遷移パターンに与えられるアークとを直感的に理解できる視覚化インタフェースについて述べる．すなわち，図 2 と同じ形式の二次元インタフェースを図 3 のように構築した．

インタフェースは，単語ラベルに対応する 6 つのボックスと，単語間の遷移を表すアークに対応する 7 本の線から構成される．すなわち，真ん中の川がテキストの主題に関係する本流を表し，左上に主題に未関連な水源を，右下が本流から分岐して流れる支流を表している．

インタフェース上に表示する単語は，テキストの最初から任意の途中のセグメントまでに出現した全単語であり，指定するセグメント終了時点で各単語に与えられているラベルに対応するボックスに，各単語が収められる．

表 1: 実験に用いたテキスト

| ジャンル | TEXT | SEG | LABEL | LABEL /SEG |
|---|---|---|---|---|
| 論文の原稿 | 5 | 59 | 1818 | 31 |
| 昔話 | 5 | 31 | 191 | 6 |
| ニュース記事 | 10 | 10 | 165 | 16 |
| コラム記事 | 5 | 9 | 118 | 13 |
| ブログエントリ | 11 | 20 | 149 | 8 |
| ブログコメント | 5 | 100 | 954 | 10 |
| 掲示板 | 8 | 74 | 691 | 9 |

またこのインタフェースでは，各セグメント終了時点の状態（各単語のラベルとその位置）を 1 シーンとして，連続するシーンを滑らかに補間する（単語がボックス間を移動する）アニメーションを再生することができる．これによって，テキストの最初から最後までの単語の出現状況や，ラベルの変化を続けて見ることができる．

また，単語の流れの理解を助けるため，以下の工夫を施している．

1. 各単語は各ボックス内で，テキスト内での単語の出現順に表示される．
2. 各単語は，そのラベルに対応した色がつけられる．ただし，NEW,BYWAY を経由した単語は，NEW,MAIN のボックス内ではその色を維持する．一度 SUB になった単語は，以降ずっと SUB の色を維持する．
3. 単語の出現頻度が大きくなるにつれ，フォントサイズが大きくなる．
4. アニメーション時に，ボックス間を移動する単語数に応じてアークが太くなる．
5. 過去 10 セグメント現れなかった単語は，徐々に薄くなっていく．
6. 入力テキスト中で，アーク Convergence の遷移を行う単語には下線を引く．

1. はテキストのどの辺りで出てきた単語なのか，また単語の出現順序関係の情報を与えるためである．2. の NEW,BYWAY の色に関しては，もともと主題に関係する単語だったのか，後から主題との関係を与えられた単語なのかを区別するため，SUB の色は一度深く掘り下げられた話題に関係する単語であることを明示するために設けた．3. は単語の出現頻度による重要度を表すためであり，4. は各セグメント間での単語の移動を要約する目的で設けた．5. は全ての単語を表示することで煩雑になることを避けるとともに，各セグメント終了時点におけるテキストの読み手の思考状態に近いイメージを作成することを意図している．6. はテーマの結論に結びつく重要語に線を引くことで，どのタイミングでどこに重要語が出現したかを確認するための情報として用いる．

## 4. ラベル付けアルゴリズムの妥当性評価

テキスト中の各単語に与えられるラベル，およびラベルの変化を表すアークの妥当性を評価する実験を行なった．用意したテキストは，表 1 に示す，論文の原稿，昔話，Web 上の

表 2: テキスト中の単語に各ラベルが与えられた割合 (%)

| ジャンル | TOP | SEED | MAIN | SUB | NEW | BYW |
|---|---|---|---|---|---|---|
| 論文の原稿 | 22 | 15 | 51 | 4 | 6 | 2 |
| 昔話 | 28 | 23 | 8 | 4 | 31 | 7 |
| ニュース記事 | 20 | 41 | 13 | 4 | 19 | 2 |
| コラム記事 | 20 | 41 | 12 | 4 | 21 | 2 |
| ブログエントリ | 19 | 30 | 8 | 5 | 32 | 6 |
| ブログコメント | 17 | 23 | 17 | 11 | 24 | 8 |
| 掲示板 | 17 | 22 | 14 | 8 | 27 | 12 |

ニュース記事とコラム記事，ブログ (日記)，ブログコメント，掲示板の計 49 テキストである*3．ただし，表中の TEXT は各ジャンルのテキスト数，SEG はセグメント数，LABEL はラベルが付与された単語の総数，LABEL/SEG は 1 セグメント当たりのラベルが付与された単語数を表し，各数値は各ジャンル内での平均となっている．またセグメントの与え方は，ブログコメントと掲示板のテキストに関しては，1 つのコメントや書き込みを 1 つのセグメントとし，それ以外のテキストでは，空行や段落の切れ目をセグメントの区切りとした．

表 2 に，テキスト中の単語に，各ラベルが与えられた割合（各ジャンル内での平均値）を示す．

論文の原稿では，MAIN の割合が高く，主題に沿った多くの説明がなされていたことが伺える．昔話では，TOPIC の値が高い反面，MAIN の割合が低くなっており，物語の主人公がさまざまに場面が変化する中で話が展開したことに対応すると考えられる．ニュースやコラム記事では，主題に関係する新しい単語である SEED の割合が高く，幅広い情報提供の意味合いが強かったことが伺える．ブログエントリでは，MAIN の割合が低く，主題に沿った一貫した話ではなく，ブログの作者が思いつきのままに，さまざまなことを書いていたと考えられる．ブログコメントと掲示板では，BYWAY の割合が高く，主題とは関係のない話題についての話が続きやすかったことが伺える．その他，ブログエントリ，昔話，掲示板の順に主題に未関連の新出単語 NEW の割合が高くなった理由は，これらのテキストでは，話の流れが読めず，全く予想外の方向に話が進むことが多かったためと考えられる．また，ブログコメントや掲示板において，単語が関連と非関連の間を行き来するラベル SUB の割合が高くなっており，これは対象としたテキストにおいて，類似の議論が繰り返し起こり議論のループが見られたことが原因と考えられる．

表 2 における，TOPIC 関連のラベル (左の 4 つ) を含む割合は，論文の原稿，ニュースとコラム，ブログコメントと掲示板と昔話，そしてブログエントリ，の順に多くなった．すなわち，論文の原稿には主題に関係する単語が繰り返し出現するため，主題との関係が強いと判断され，ニュースやコラム記事については，主題に関係して短く完結にまとめられている．昔話は主題に関する一貫性があると予想される反面，さまざまな伏線や背景描写の記述も含まれるため，全体としての TOPIC 関連度は下がる．掲示板やブログのコメント欄については，その掲示板の主題や対象となるブログエントリの制約の下で，比較的自由な記述が可能であったため，TOPIC との一定の関連度を保ちつつも，特に高い値にはなっていない．ブログエントリは，書き手が何の制約も受けずに記述することが可能なため，最も主題に関する揺れが生じたと考えられる．以上のことから，システムのラベルづけによって表される各テキストの主題との関連度は妥当な数値だったと言える．

*3 いずれも，テキストの流れを把握することを目的としているため，極端に短いテキストは含んでいない．また，ブログコメントと掲示板においては，一部の人々の間で話題になるなど，盛んに書き込みが行われていたブログや掲示板の中から，最初の 100 コメントを抜き出して用意した．

## 5. 視覚化インタフェースの使用方法

本章では，テキスト中の各単語に与えられるラベルおよびアークをもとに，テキストの話の流れを直感的に把握するために，視覚化インタフェースを用いる方法を示す．

テキスト中の各単語に与えられるラベルの量が，視覚化インタフェース上の川幅を表現し，アークがラベルの変化、すなわち川の流れの変化としてアニメーションで表される．すなわち，視覚化インタフェースにおける見るべきポイントとその意味は下記で表される．

1. 本流の川幅：テキストの主題に関係する単語の数
2. 支流の川幅：テキストの主題に関係する話の中で，特に掘り下げられた話に関係する単語の数
3. 水源の水量：テキストの主題に未関連な話のもとになりえる単語の数
4. 水源の水量や川幅の変化を与えるタイミング：テキストの話の流れに影響を与えるセグメント
5. 表示される単語の濃さや大きさと下線の有無：テキスト中の各時点で重要な位置付けにある単語

## 6. 結論

本稿では，テキスト中の各単語に，テキストの主題との関係を表すラベルを付与し，テキストの話の流れを明らかにする川下りシステムとその視覚化インタフェースについて述べた．本システムがテキスト中の各単語に与えるラベルの妥当性を検証し，テキストが主題に関して一貫性があるかを判断できることを確認した．

本システムは，一度に読み切ることができない長いテキストや，コメントの多い電子掲示板などに対して用いることで，その全体像を把握することに役立てられると期待できる．また，論文などのテキストを作成する際に，主題に関する一貫性の有無や，各単語と主題との関わり方を確認することで，テキストの推敲支援にも用いることが可能と考えている．

## 参考文献

[1] 相良直樹・砂山渡・谷内田正彦：サブトピックを考慮した重要文抽出による報知的要約生成, 電子情報通信学会論文誌, Vol.J90-D, No.2, pp.427 – 440, (2007).

[2] Wataru Sunayama and Masahiko Yachida: Panoramic View System for Extracting Key Sentences Based on Viewpoints and an Application to a Search Engine, Journal of Network and Computer Applications, Elsevier Science, Netherlands, Vol.28, No.2, pp.115 – 127, (2005).

[3] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸: 形態素解析システム『茶筌』, Version 2.2.9，使用説明書, (2002).