原著論文

Original Paper

# 光と影を用いたテキストのテーマ関連度の可視化

## Text Visualization using Light and Shadow based on Topic Relevance

西原 陽子
Yoko Nishihara
東京大学大学院工学系研究科
School of Engineering, The University of Tokyo
nishihara@sys.t.u-tokyo.ac.jp

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

砂山 渡
Wataru Sunayama
（同　上）

**keywords:** text visualization, light and shadow, information comprehension, topic relevance

**Summary**

There are so many opportunities to transmit text information on the Web. Since texts on the Web are not always written by professional writers, those may not be coherent or may be hard to be comprehended. Therefore, we should take too much time and energy to grasp topic relevance of a text.

This paper describes HINATA system that visualizes texts using light and shadow based on topic relevance. Topic is defined as a set of words such as nouns contained in a title of a text. The light expresses sentences related to a topic, and the shadow expresses sentences unrelated to a topic. This visualization method efficiently supports users for finding the parts related to a topic, and for grasping relations between sentences of a text and a topic. Experimental results showed that the proposed system could support users for understanding how a text was related to a topic.

## 1. は じ め に

近年，Web の発達により，個人が簡単に情報を発信，取得できるようになってきた．特に，メール，Web サイト，blog，電子掲示板を通じて，文章の交換をすることが，ごく当たり前になって来ている．最近では，個人の情報発信の手段として，YouTube[*1]やニコニコ動画[*2]など，動画をメインのコンテンツとした Web サイトも注目されている．しかし，これらのサイトにおいても，動画へのコメント機能を設けるなど，コミュニケーションの手段として文章を扱う点は外すことができない．

Web 上に存在する文章の多くは，文章を書くプロではないごく普通の個人が書いたものであるため，必ずしも，文章の主題と内容の一貫性が保たれているわけではない．そのため，理解が難しい文章や，意図がつかみづらい文章，無駄に長い文章等が存在している．またプロが書いた文章であっても，長い文章を読むには労力がかかったり，読み手によって，読みたい部分が変わる可能性がある．

そこで本研究では，文章の主題 (テーマ) と各文の関連度を評価し，文章中の各文を光と影により可視化する，ひなたシステムを提案する．文章のテーマと関係のある部分，および関係のない部分を可視化することで，テーマと関係のある箇所と，その全体に対する位置や割合を確認でき，テーマとの関係に基づいて文章の理解を支援できると考えられる．また本研究では，人間が文章を先頭から読み進める過程における，テーマと文章との関連度を光と影として表す．

以下，本論文では，2 章で関連研究, 3 章でテキストに光と影を与えるひなたシステムの構成と詳細について述べる．4 章で本システムの光と影が，テキストの内容把握に対して有効に働くことを確認した評価実験について述べ，5 章で本論文を締めくくる．

## 2. 文章理解支援の関連研究

本章では文章理解支援の関連研究について述べる．

### 2·1 文章の表示方法と理解の関連

文章の表示方法を工夫すると，その内容を理解しやすくなることが，従来研究において明らかにされてきた．文章を読む際には，自らが重要と思う箇所に線を引くと文章の理解が進むことに加え，他人によって下線が引かれた箇所に注意して読むことでも文章の理解につながることが知られている [魚崎 00]．また，重要ではない文を

*1 http://www.youtube.com/
*2 http://www.nicovideo.jp/

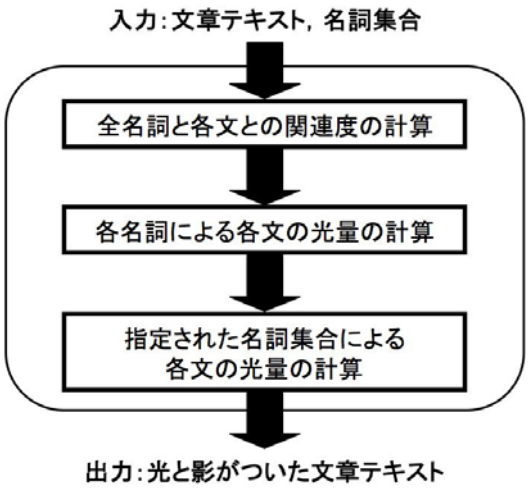


**図 1** ひなたシステムの構成

削除して表示すること [Radev 02] なども効果があると知られている．これらの研究に対し，本論文では文章に背景色をつけることによる，文章理解への効果を明らかにする．

これまでの研究でも文章に背景色をつけることにより，文脈変化を理解させる効果があることは知られている [Smith 01]．これに対し本研究では，文章のテーマとの関連を背景色により表現することで，文章のテーマとの関連，内容把握を支援する．

### 2·2 文章構成の評価と可視化

文章の構成を評価して可視化するシステムが多数提案されてきており，議論の議事録の構成を可視化し，議論支援を行うシステムがある [藤井 08, Hotta 00, 松村 03, 西田 06]．例えば松村らのシステムでは，議事録を話題ごとに分割し，分割された議事録間の関連を評価して，議事録の構造を可視化することができる [松村 03]．また，小論文のように一人の人間によって書かれた比較的長い文章に対しても，複数のシステムが提案されてきている [Barzilay 08, Burstein 04, 石岡 08]．例えば石岡の手法では，小論文中の接続詞と文末表現を用いて小論文の構成を評価し，評価結果を星座盤により可視化する [石岡 08]．これらの研究と本研究は，文章構成の可視化という点で共通しているが，本研究では文章と文章の関連ではなく，テーマと文章の関連を評価する点が異なる．すなわち，文章のテーマとの関連を評価することにより，テーマに関しての文章理解を支援する．

テーマと文章の関連を評価するシステムに KeyGraph がある [大澤 99]．KeyGraph は文章中の単語の関係を，単語をノード，関係をエッジとしたグラフにより表現する．KeyGraph により，テーマと文章全体の関係は理解できるが，テーマと文章中の各文との関係を理解することは難しい．本研究では，各文とテーマの関係を評価し，その関係が理解可能な可視化を行い，文章理解を支援する．

文章の構成ごとに異なる色を用いて文章の背景に彩色し，文章の理解を支援するシステムが提案されている [内田 97]．このシステムで彩色を行うのはユーザであり，ユーザの主観に基づき彩色を行うことが可能になっている．この研究と本研究とでは，文章の背景に彩色して文章理解を支援する点で共通しているが，本研究ではシステムが自動的に彩色を行い，またユーザは彩色パターンを決める単語を変更できる点でこの研究とは異なっている．すなわち，システムが半自動的に彩色を行うことにより，彩色のための労力を軽減できると考えられる．

### 2·3 文章の要約

従来からの文章要約研究も，文章のテーマに関する理解を支援する．文章要約手法の多くは文章中の各文に対して重要度を与え，重要な順に文を抽出して出力する．重要度の評価方法としては，高頻度の単語が多数含まれるほど重要とするものや，テーマに関連する単語を多数含むほど重要とするものなどがある [Knight 02, 大竹 02, 相良 07, 砂山 01]．しかし要約システムの場合，要約として選ばれなかった文は出力に含まれることはない．提案するシステムでは，テーマとの関連が強い文ほど重要な文として光を当てて出力するが，重要でない文にも暗い光や，影を当てて出力を行なう．そのため，要約システムでは見落とされがちな文を確認できたり，重要な文のテキスト全体の中での位置を直感的に把握することが可能になると考えられる．

## 3. ひなたシステム

本章では，文章テキストと，名詞の集合（ユーザが考える文章のテーマ）を入力として，テーマに関係する部分を光，関係しない部分を影で表現する，ひなたシステム (図 1) について述べる．

### 3·1 入力：文章テキストと名詞集合

本システムは，可視化を行いたいテキストと，そのテキストに光を与えるもととなる名詞の集合を入力とする．

対象となる入力テキストは，句点等によって一文ごとに切り分けが可能なテキストで，かつ形態素解析によって名詞の抽出が可能なテキストとする．なお形態素解析には,「茶筌」[松本 02] を用いた．

加えて，ユーザが考える文章のテーマを名詞の集合として与える．ユーザがテーマを指定せず，かつテキストのタイトルが存在するときには，そのタイトルに含まれる名詞集合を入力とすることもできる．また，一度可視化の結果を表示させた後に，再度，初めに与えた名詞集合とは異なる集合を与え，インタラクティブな可視化を行うことも可能である．

**表 1** 各文の関連度に対する背景の RGB 値

| 関連度 | R 値 | G 値 | B 値 |
|---|---|---|---|
| 4.0 以上 | 255 | 255 | 55 |
| 3.0 以上 4.0 未満 | 220 | 220 | 0 |
| 2.0 以上 3.0 未満 | 190 | 190 | 0 |
| 1.0 以上 2.0 未満 | 120 | 120 | 0 |
| 0.1 以上 1.0 未満 | 100 | 100 | 0 |
| 0 | 0 | 0 | 0 |

### 3·2 全名詞と各文との関連度の計算

本節では，テキスト中の各文に，各名詞との関連度を与える方法について述べる．本モジュールでは，以下で述べる方法を，テキスト中の全名詞[*3]に対して適用する[*4]．

ある名詞 $w$ に対する，$i$ 番目の文の関連度 $R_i(w)$ は，その文に含まれる，名詞 $w$，およびその名詞の関連語の数として与える．具体的には，名詞 $w$ の関連語集合 $Rset(w)$ を元に，以下のアルゴリズムによって計算する．

**[関連度計算アルゴリズム]**

1. $i=1, Rset(w)=\{\}$ と初期化する．
2. $i$ 番目の文の関連度を，式 (1) で与える．ただし，$F_i(w)$ は，名詞 $w$ の $i$ 番目の文における出現頻度，$\alpha$ は関連語の重みを表すパラメータ[*5]とする．

$$R_i(w) = F_i(w) + \alpha \times \sum_{w' \in Rset(w)} F_i(w') \tag{1}$$

3. $i$ 番目の文に，名詞 $w$ が含まれていれば，その文に含まれているその他の名詞をすべて，$Rset(w)$ に追加する．
4. テキストの最後の文に達していれば終了．そうでなければ，$i=i+1$ として，2 へ．

すなわち，指定された名詞と一文中で共に使われた名詞を，関連語に加えながら，各文の関連度を与えていく．これは，テキストを前から順に読み進める際に，そこまでの内容から，単語間の関係が理解できるかどうかを表す指標として，そこまでに同じ文中で使われた単語同士であれば関係が理解できる，との考えに基づいている．本アルゴリズムでは，テキストの後半になるほど，$Rset$ に含まれる単語の数が増え，後半の文ほど関連度が高くなる傾向がある．これは本システムで可視化する光と影が，人間がテキストを読み進める際に，各文と与えられたテーマとの関連を判断できるか否かを表したことによる．すなわち我々がテキストを読む際に，読み初めの時点では何についての話なのかを読み取る必要があり，分かりにくい部分もあるが，読み進めるうちに，話の流れや単語間の関係をつかんで理解しやすくなっていくことに準じている．

### 3·3 各名詞による各文の光量の計算

本節では，各名詞の各文との関連度をもとに，各文を照らす光の量を計算する方法について述べる．

光は，指定された名詞の関連度をもとに，テキストを表示する際の各文の背景色として表 1 の 256 段階の RGB 値[*6] を与える．すなわち，赤 (R) と緑 (G) と青 (B) の組み合わせによって黄色の光を作成する．また関連度が 0 の時には，背景が真っ黒になり，テキストも黒色で表示される場合，テキストは全く見えなくなる．

### 3·4 名詞集合による各文の光量の計算

名詞集合 $W$ による $i$ 番目の文の光量は，集合内の各名詞の $i$ 番目の文の関連度を合計した式 (2) の関連度 $L_i$ を，表 1 に適用した値とする．

$$L_i = \sum_{w \in W} R_i(w) \tag{2}$$

すなわち，名詞集合中の 1 つの名詞と強く関係している文や，名詞集合中の多くの名詞と関係している文に，大きな光量が与えられる．

### 3·5 出　　力

本システムの出力インタフェースを図 2 に示す．インタフェースの右側に，光と影をつけたい文章テキストを，コピー&ペーストで貼付けて入力を与える．インタフェース左下の「解析」ボタンを押すと本システムの処理が実行され，光と影をつけたテキストを，インタフェース左側に出力する．図 3 に，テキストの可視化例を示す．光と影をつける名詞集合として，タイトルが指定されている場合には，タイトル中の名詞集合を用い，タイトルがないテキストの場合には，テキスト全体によく光を当てる，式 (3) の値が高い名詞 5 つを用いる．ただし式中の $n$ は，テキストの文の数とする．

$$text(w) = \sum_{i=1}^{n} R_i(w) \tag{3}$$

またインタフェース下部には，光を与える名詞を変更するための，名詞リストとチェックボックスがある．左から,「タイトル中の名詞」「式 (3) の関連度が高い名詞」「出現頻度の高い名詞」「多くの文を光らせる名詞の組合せ」

---

*3 名詞の中でも，形態素解析により，一般，固有名詞，サ変接続，形容動詞語幹，副詞可，接尾-助数詞に分類される単語を対象とした．

*4 名詞のみを対象とした理由は，テキストにおいて使用される単語の約 7 割が名詞というデータ [砂山 06]（評価実験で使用したテキストにおいても当てはまった）と，多くのテキストの主題は名詞で表現されることが多いと考えたことによる．ただし,「食べる」などの動詞や他の品詞の単語がテキストの主題となることも考えられ，テキストに応じて他の品詞を対象に加えることも可能となっている．

*5 実験的に $\alpha=0.5$ としている．

*6 RGB 値とディスプレイ上の輝度は非線形の関係にあるため，見た目の輝度差を元に調整した値となっている．

**図 2** インタフェースの概観

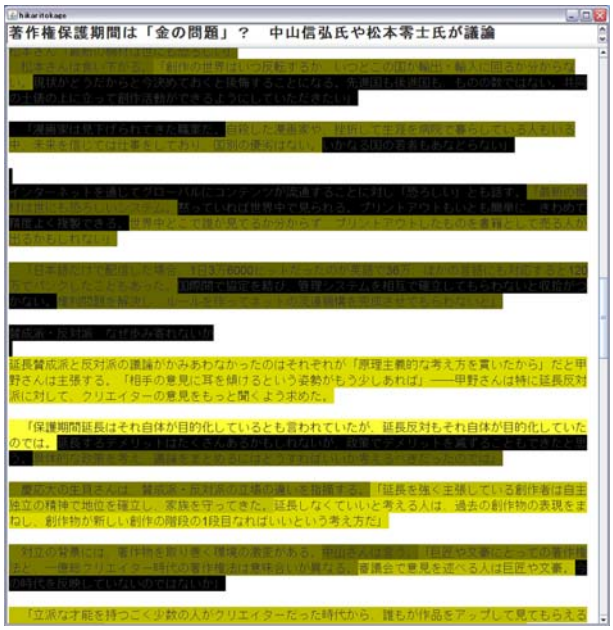


**図 3** テキストの可視化例

「現在チェックされている名詞」の５つのリストを表示している．チェックボックスを用いて名詞を選択することにより，光を与える名詞を変更することができる．ユーザはこの名詞リストを用いて，選択する名詞とテキストとの関係を視覚的な光と影によって捉え，各名詞とテキストとの関係の理解に役立てる．

なお,「多くの文を光らせる名詞の組合せ」には，次の Greedy アルゴリズムを用いている．

1. 名詞集合 $C = \{\}$ とする．
2. 名詞集合 $C$ によって光が与えられる文 (R 値と G 値が 0 でない文) に比べ，集合 $C \cup \{w\}$ によって，光が与えられる文の数が最も増える名詞 $w$ を，集合 $C$ に加える．

   光が与えられる文が増える名詞がなければ終了．
3. 名詞集合 $C$ によって光が与えられる文の数が，テキストの文の総数の 9 割以上となれば終了．そうでなければ，2 へ．

### 3·6 想定する本システムの使用方法

想定する本システムの使用方法の主なものとして，文章理解支援，および文章のタイトル作成支援が挙げられる．

#### §1 文章理解支援

テキスト中の，自分の興味に合う部分だけを素早く探して読みたいユーザを仮定する．

テキストを入力としてシステムに与えると，ユーザは光が当たっている明るい部分だけに着目し，暗い部分を飛ばして読むことができ，内容のすばやい把握が期待できる．またユーザは，再度自分の興味ある名詞を名詞リストから選んで指定することで，自分の興味に関連するテキスト部分を，優先的に探して読むことができる．

また光と影が与えられたテキスト全体をスクロールして見渡すことで，テキストの何割程度に光が当たっているか，またテキスト中のどの部分に特に光が当たっているかを確認して，自分の興味とテキストとの関係を理解することができる．

#### §2 文章のタイトル作成支援

あるテキストに適切なタイトルを付けたいユーザがいると仮定する．

タイトルは，テキスト全体をよく表す表現にしたいと考えるため，テキスト全体に光を与えられる名詞は参考になると考えられる．すなわち，テキスト全体に光が当たるような名詞の集合を，自分の好みに応じて選んでいくことで，タイトルに含めるべき名詞の集合を用意することができるようになると考えられる．

一方で，本文中にない単語を用いて，タイトルを作成したい場合もある．その際には，まず提案システムにより文章との関連が強い単語を取得し，それらの単語から連想される他の単語やフレーズなどを考えてもらう支援が可能と考えられる．

## 4. 評 価 実 験

本章では，提案するひなたシステムが出力する，光と影の有効性を確認するために行った，光量の妥当性評価，文章理解支援，文章タイトル作成支援の３つの実験について述べる．

### 4·1 光量の妥当性の評価

#### §1 実験内容

表 2 の 4 つのカテゴリの合計 16 テキスト (各カテゴリ 4 テキスト) に対して，被験者に各テキストのタイトルと関連のある文を，テキスト中から選んでもらい，システムの出力する光と影が与えられる文と比較する実験を行なった．各カテゴリのテキストの特徴について，ニュース

**表 2** 実験に用いたテキストとその特徴

| カテゴリ | 一貫性 | 話題数 | 話の流れ |
|---|---|---|---|
| ニュース | 高い | 1つ | あり |
| ブログ | 低い | 複数 | あり |
| レビュー | 高い | 1つ | なし |
| 2ちゃんねる | 低い | 複数 | なし |

**表 3** 被験者が回答した関連文の数

| カテゴリ | 関連文の数 (割合) | 全文数 |
|---|---|---|
| ニュース | 109 (83%) | 131 |
| ブログ | 40 (46%) | 87 |
| レビュー | 107 (68%) | 158 |
| 2ちゃんねる | 298 (40%) | 746 |

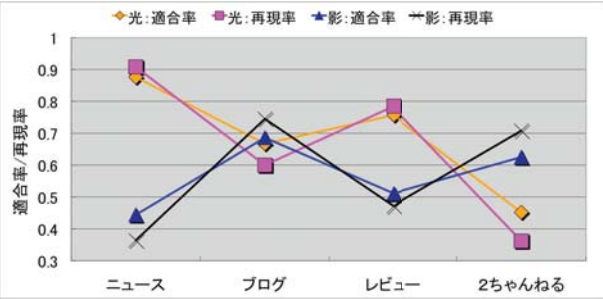


**図 4** システムの光と影の適合率と再現率

とレビュー[*7]は明確な1つのものについて述べられているので，一貫性が高く話題が1つのテキストとした．話の流れの有無は，ニュースとブログは一人の人が書いているのに対して，レビューと2ちゃんねる[*8]は複数人が書いたテキストとなっていることを元にした．

被験者は情報科学を専攻する大学生，大学院生 28 名で，テキストを前から順に一文ずつ，後戻りしないで読んでもらったときに，全ての文に，タイトルに関係がある，ない，どちらともいえないの3択で回答してもらった．一人の被験者には 4 つのテキスト (各カテゴリ 1 テキストずつ) を評価してもらったため，1テキスト当たりの回答数は 7 名分となった．

### §2 実験結果

過半数の被験者が，タイトルと関係があると回答した関連文の数を表 3 に，システムの光 (タイトルに含まれる名詞集合と関連する文) と影 (光以外の文) の適合率と再現率を図 4 に示す．ただしシステムの光としての出力は，式 (2) によって 0.5 以上の関連度が与えられた文，影としての出力は，その関連度が 0 となった文とした．

### §3 考 察

図 4 のシステムが出力する光の適合率と再現率について，1つの話題について一貫性がある，ニュース，レビューにおいて高い値となった．これは，本システムが同じ文中に存在する名詞を手がかりとして関連度を与えているため，一貫性があるテキストにおいては似た単語が使われやすく，関連がある文をうまく抽出できたことによると考えられる．

一方，ブログ，2ちゃんねるにおいては，表 3 において関連する文の割合がともに低くなっており，光の適合率，再現率は，ニュース，レビューほど高い値とはならなかった．特に2ちゃんねるにおいては，複数の人で雑談をすることが多く，前の書き込みを反映して議論が発展する傾向が低かったことにより，最も値が低くなったと考えられる．

しかし，ブログ，2ちゃんねるにおいては，図 4 の影の適合率と再現率が，ニュース，レビューに比べて高い値となった．これは，ニュースやレビューにおいて，タイトルとの関係がやや乏しい前置きや，例示にも，タイトル中の名詞が含まれていたことが原因として挙げられる．逆に，ブログや2ちゃんねるにおいては一文が短く，タイトル中の名詞がない文が積極的に除かれたことによって，値が高くなったと考えられる．

以上より，本システムは，ニュースやレビューのように一貫性の高いテキストについては光を，一貫性が低いテキストについては影の出力を参考にすることができ，テキストの話の一貫性の有無によらず，光と影の出力による一定の効果が期待できることがわかった．

**表 4** 文章理解実験に用いたテキストと文数

| タイトル | 文数 |
|---|---|
| A:衝撃的事実　士農工商はなかった武士は誰でもなれた | 1947 |
| B:普通の電池は充電しても使える | 1342 |
| C:良い電子辞書教えて！ | 1300 |
| D:ネットから長文が消えたいくつかの理由 | 71 |
| E:データは消えない――メモリカードや USB メモリに潜む落とし穴 | 52 |
| F:著作権保護期間は「金の問題」？　中山信弘や松本零士氏が議論 | 164 |

## 4·2 文章理解支援実験

### §1 実験内容

情報科学を専攻する大学生，大学院生 20 名に，テキスト中から，テキストのタイトルに関連して重要であると思う文を 5 文抜き出してもらった．重要な文を抜き出すためには，テキストの内容を理解して要点を捉える必要があると考え，そのためにかかる時間と抜き出された文をもとに評価した．実験には表 4 のタイトルのテキスト (A,B,C は2ちゃんねる，D,E,F は 3000 から 5000 字の，あるテーマに関して説明が書かれたコラム) を用いた．各テキストは表 5 に示す通り，A,B,C,D のテキストは影の表示が多く，E,F のテキストは光の表示が多くなっていた．制限時間を 5 分とし，制限時間を超えても 5 文抜き出せてない場合は，さらに時間を延長して，5 文抜き出せるまでの時間を計測した．

各被験者には，以下の提案システムと比較システムとを用いてもらい，6 つのテキスト全てについて，指定さ

*7 http://kakaku.com/
*8 http://www.2ch.net/

**表 5** 各テキストにおける，文の関連度の分布
(下線は上位 20%の境界を含む階級)

| 関連度＼テキスト | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 4.0 以上 | 4 | 125 | 49 | 0 | 15 | 22 |
| 3.0 以上 4.0 未満 | 28 | 128 | 82 | 0 | 13 | 15 |
| 2.0 以上 3.0 未満 | 138 | 199 | 195 | 9 | 12 | 32 |
| 1.0 以上 2.0 未満 | 376 | 411 | 278 | 28 | 11 | 61 |
| 0.1 以上 1.0 未満 | 613 | 480 | 697 | 34 | 2 | 35 |
| 0 | 788 | 994 | 916 | 47 | 4 | 29 |

**表 6** 5 文抜き出す時間が 5 分を超えた人数 (10 人中)

| システム＼テキスト | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 提案システム | 4 | 4 | 1 | 2 | 2 | 1 |
| 要約システム | 7 | 7 | 6 | 2 | 0 | 5 |
| 全文システム | 10 | 10 | 10 | 6 | 5 | 10 |

**表 7** 2 人以上が抜き出した文の数

| システム＼テキスト | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 提案システム | 10 | 8 | 12 | 10 | 12 | 10 |
| 要約システム | 11 | 8 | 11 | 9 | 10 | 13 |
| 全文システム | 6 | 7 | 7 | 11 | 9 | 9 |
| 提案と要約で共通 | 3 | 2 | 4 | 5 | 7 | 9 |
| 提案と全文で共通 | 6 | 6 | 6 | 9 | 7 | 8 |
| 要約と全文で共通 | 2 | 2 | 1 | 3 | 3 | 2 |

**表 8** 各システムを用いた被験者が抜き出した関連度 (式 (2))3 以上の文の数と，その適合率と再現率

| システム＼テキスト | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 提案システム | 21 | 42 | 38 | 0 | 50 | 41 |
| 適合率 | 0.42 | 0.84 | 0.76 | – | 1.00 | 0.82 |
| 再現率 | 0.25 | 0.10 | 0.15 | – | 0.68 | 0.49 |
| 要約システム | 3 | 29 | 4 | 0 | 50 | 49 |
| 適合率 | 0.06 | 0.58 | 0.08 | – | 1 | 0.98 |
| 再現率 | 0.06 | 0.08 | 0.02 | – | 0.36 | 0.57 |
| 全文システム | 10 | 21 | 16 | 0 | 35 | 31 |
| 適合率 | 0.2 | 0.42 | 0.32 | – | 0.7 | 0.62 |
| 再現率 | 0.16 | 0.04 | 0.05 | – | 0.43 | 0.27 |

れたいずれかのシステムを用いて回答してもらった．

- 提案システム：テキスト全文を表示する．各文にテキストのタイトル中の名詞との関連度 (式 (2)) にもとづいて，光と影を出力する．
- 要約システム：テキスト全体の 20%の文を表示する．提案システムが文に与える，式 (2) の評価値が高い文のみを，出現順に並べて表示する．
- 全文システム：テキスト全文を表示する．

### §2 実験結果

被験者が 5 文抜き出すまでにかかった時間が 5 分を超えた人数を表 6 に示す．また，2 人以上が抜き出した文の数を表 7 に，関連度 (式 (2))3 以上の抜き出された文の数と，抜き出された文の内，関連度 3 以上の文に限定して算出した適合率と再現率を表 8 に示す．

### §3 考 察

表 6 から，A,B,C,F の 4 つのテキストについて，提案システムは他のシステムに比べて，5 文抽出を 5 分以内で終えられる人数が 3 人から 9 人多くなった．これは，提案システムにおいては，要約システムの 5 倍の量のテキストを表示しているにもかかわらず，ユーザは光と影を参考にすることで，内容把握ができたためと考えられる．

D,E の 2 つのテキストにおいては，提案システムと要約システム，いずれのシステムにおいても，ほぼ 5 分以内で抽出できる結果となった．これは，この両テキストが最も長いテキスト A に対し，それぞれ 27%，37%の文数であり，要約システムにおいてはスクロールなしで 1 画面ですべてのテキストを閲覧できる状態であったためと考えられる．

このことから，画面スクロールが必要な量のテキストの内容把握においては，本システムが出力する光と影によって，内容把握が支援できると考えられる．

表 7 において，提示する文の数が多かった提案システムにおいて，抽出された文数の分布が要約システム，全文システムと同程度となった．また，関連度 3 以上の文が存在しなかったテキスト D を除いて，共通に抽出された文は全て関連度が 3 以上の文であった．このことから，光と影の出力によって，複数の人が重要と思える客観的に重要な文を，多くの文の中から選べていることがわかる．さらに，異なるシステム間で共通に抽出された文は，提案システムと全文システムが最も多かった．このことから，提案システムにより全文システムと同程度の内容把握が行えたと考えられる．

また表 8 から，2 ちゃんねるのテキスト A,B,C においては，提案システムでタイトルとの関連度が 3 以上の文が要約システムよりも多く選ばれる結果となった．このことから，長く一貫性のないテキストにおいては，重要な文を選ぶための指針がないと，うまく重要な文を選べない可能性が高いことがわかり，本システムの出力する光と影は，そのようなテキストの理解にも役立てられると考えられる．さらに表 8 から，提案システムの方が，全文システムよりも関連度 3 以上の文を 10 文以上多く抽出する結果となった．また，テキスト A,B,C,E では、関連度 3 以上の文の適合率と再現率は，提案システムが最も高くなった．このことから，提案システムにより抽出される文は妥当なものが多く含まれることが分かった．

## 4·3 タイトル作成支援実験

### §1 実験内容

情報科学を専攻する大学生，大学院生 20 名に，あるテキストのタイトルに使用すべきだと思う名詞を，5 個以上 10 個以下で選んでもらい，選んだ名詞の出来るだけ多くを使ってテキストの内容を良く表すタイトルを作成してもらった．タイトル作成にかかった時間，選ばれた名詞，作成されたタイトルによって評価を行なった．

**図 5** タイトル作成支援実験に用いたインタフェースの概観
名詞システムでは，左上の可視化領域は使用しない

**表 9** タイトル作成実験に用いたテキストのタイトルと文数

| タイトル | 文数 |
|---|---|
| G:医療の「テレビドラマ」は増えてます．でも「報道」が日本医療をダメにしている？ | 48 |
| H:日本の男はなぜ勝負に弱いのか　五輪野球･サッカーで考えた | 78 |
| I:「グレシャムの法則」から見た基軸通貨ドルの明日 | 61 |
| J:デフレ下の販売奨励金廃止で形態メーカーは敵対的買収の餌食になる | 90 |
| K:首都圏のスーパーが電子看板で売上増　ソニーが開拓する新たな「広告メディア」 | 59 |
| L:取材の方法論を変えたハイビジョンカメラ | 115 |

タイトルを作成するテキストには，表 9 の 3000 字から 5000 字の比較的長い，あるテーマに関して説明が書かれたコラムを用いた．

各被験者には，6 つのテキスト全てについて，以下のいずれかのシステムを用いて，提示される名詞リスト横のチェックボックス (図 5 右下) を使って，タイトルにふさわしい名詞を選んでもらった上で，タイトルを作成して回答してもらった．

- 提案システム：テキスト全文と，関連度 (式 (2)) が高い順に並べられた名詞リスト，および選択された名詞集合によるテキストの光と影を表示する．
- 名詞システム：テキスト全文と，出現頻度順に並べられた名詞リストを表示する．
- 全文システム：テキスト全文を表示する．

## §2 実験結果

4 人以上の被験者がタイトルに使った名詞を表 10 に，被験者が作成したタイトルの例を表 11 に，タイトル作成までにかかった時間の中央値を表 12 に示す．またタイトル作成のために選択された名詞，および実際のタイトルに使われた名詞について，提案システムと名詞システムにおいて，他方のシステムに比べて 2 人以上多くの被験者が使用した名詞の，テキスト内での出現範囲[*9]の全テキストに対する割合 (被験者平均) を図 6 に示す．

**表 10** 4 人以上の被験者がタイトルに使った名詞
（下線は全文システムを用いた被験者が 4 人以上使った名詞）

| テキスト | | 名詞 |
|---|---|---|
| G | 提案 | 不足，医師，深刻，メディア，産婦人科 |
| | 名詞 | 医師，医療，不足，メディア |
| | 全文 | メディア，医師，影響，不足，問題，産婦人科 |
| H | 提案 | 勝負，競争，日本人 |
| | 名詞 | 勝負，教育，日本人 |
| | 全文 | 競争，教育，弱い，勝負，日本人 |
| I | 提案 | ユーロ，サンマリノ，悪貨，通貨 |
| | 名詞 | 価値，ユーロ，サンマリノ，需要，通貨 |
| | 全文 | ユーロ，悪貨，価値，駆逐，通過，良貨 |
| J | 提案 | 携帯，端末，デフレ，経済，産業，市場 |
| | 名詞 | 携帯，端末，経済，構造，日本， |
| | 全文 | メーカー，携帯，経済，電話，日本，市場，端末 |
| K | 提案 | 顧客，メディア，広告，新た |
| | 名詞 | 業界，小売 |
| | 全文 | デジタルサイネージ，メディア，業界，広告，小売 |
| L | 提案 | カメラ，機能，取材，ハイビジョン，ビデオカメラ |
| | 名詞 | カメラ，ハイビジョン，取材，ビデオカメラ |
| | 全文 | カメラ，キヤノン，ハイビジョン，メリット，取材 |

**表 11** 被験者が作成したタイトルの例

| テキスト | | タイトルの例 |
|---|---|---|
| G | 提案 | 医師不足とメディアによる事故報道 |
| | 名詞 | 医療現場の現状とメディアのとらえ方 |
| | 全文 | 医師不足による問題とメディアの影響 |
| H | 提案 | 日本人の勝負弱さと競争の重要性 |
| | 名詞 | 日本のスポーツ選手は勝負に弱い |
| | 全文 | ゆとり教育，反競争主義による日本人の勝負弱さの現状 |
| I | 提案 | 通貨の価値と流通の関係 |
| | 名詞 | サンマリノ共和国のユーロ硬貨の価値 |
| | 全文 | ドルやユーロの国際通貨などの悪貨が良貨を駆逐する現象 |
| J | 提案 | 日本の携帯電話事業 |
| | 名詞 | 日本の携帯端末産業の状況 |
| | 全文 | 携帯電話メーカーの市場競争がもたらす経済への影響 |
| K | 提案 | 商品広告コンテンツの多様化 |
| | 名詞 | デジタルサイネージによる利益 |
| | 全文 | 小売業界から見たデジタルサイネージ市場の拡大と影響 |
| L | 提案 | ハイビジョンカメラを取材に利用 |
| | 名詞 | ハイビジョンカメラのビジネスモデル |
| | 全文 | キヤノン製デジカメの取材におけるメリットとデメリット |

**表 12** タイトル作成までにかかった時間 (秒) の中央値

| システム＼テキスト | G | H | I | J | K | L |
|---|---|---|---|---|---|---|
| 提案システム | 449 | 486 | 506 | 710 | 559 | 588 |
| 名詞システム | 399 | 555 | 562 | 535 | 430 | 402 |
| 全文システム | 720 | 660 | 660 | 840 | 540 | 720 |

## §3 考　察

表 9 の実際のタイトルと，表 10 のタイトルに使われた名詞，および表 11 の被験者が作成したタイトルから推論するに，いずれのシステムについても，テキストの内容をおよそ捉えたタイトル付けが行なえていることが確認できる．

表 12 から，全てのテキストについて，提案システムの方が全文システムよりもタイトル作成にかかる時間が短くなった．これは，提案システムにおいては，選択する名詞により光と影の表示が変わるため，タイトルに適し

*9 テキスト中で最初に出現した文の番号と，最後に出現した文の番号の差

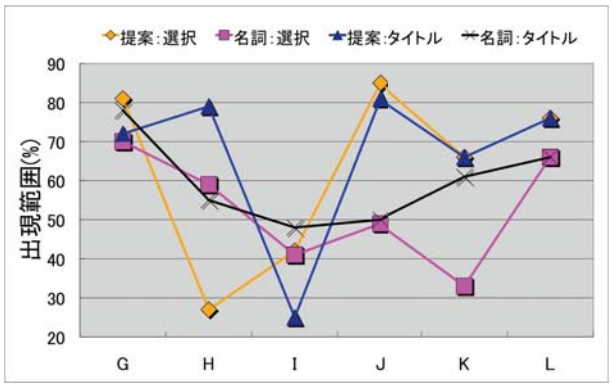


**図 6** タイトル作成のために選択された名詞，および実際のタイトルに使われた名詞について，他のシステムに比べて 2 人以上多くの被験者が使用した名詞の，テキスト内での出現範囲の全テキストに対する割合 (被験者平均)

た名詞を選択しやすかったためと考えられる．一方，表 12 から，G,J,K,L のテキストについては，提案システムの方が名詞システムよりもタイトル作成に時間がかかる結果となった．これは，提案システムにおいては，選択する名詞を変更することにより，光と影の表示が変わるため，さまざまな名詞について，光と影の当たり方を確認していたためと考えられる．

これに関連して図 6 では，提案システムにおいて，名詞選択の段階でテキスト H,I 以外，最終的なタイトル中の名詞ではテキスト I 以外において，テキストの 6 割以上の範囲で出現する名詞が，名詞システムに比べて積極的に用いられていた．このことから，ひなたシステムは，テキスト全体をよく表す名詞を使ったタイトル作成を支援できたと言える．

I のテキストは「『グレシャムの法則』から見た基軸通貨ドルの明日」という実際のタイトルに対して，「悪貨は良貨を駆逐する」というグレシャムの法則に関連した「悪貨」「良貨」という単語が，テキストの後半になって初めて使われていたため，システム間で出現範囲の差が出なかったと考えられる．しかし，提案システムでは「悪貨」を 4 人「良貨」を 3 人が用いており (比較システム 1 では各 1 人)，テキスト中の重要な部分を照らす単語の選択を支援できたと考えられる．

表 10 の下線で表された，全文システムを用いた被験者が 4 人以上使った名詞の多くは，提案システム，名詞システムを用いた被験者においてもよく使われていたことがわかる．また，提案システムと全文システムでは，選択された名詞，タイトルに使用された名詞について，他方のシステムに比べて 2 人以上多くの被験者が使用した名詞は存在しなかった．このことから，提案システムにより，テキスト全文を読み得られるタイトルと同程度のタイトル作成を支援できたと考えられる．

## 5. 結　　論

本論文では，ユーザが考えるテーマ (名詞集合) に関係するテキスト中の文を，光と影を用いて可視化するひなたシステムを提案した．本システムが文章の理解支援に有効であること，またテキストのタイトル作成に有効であることを，実験により検証した．

今後は提案システムを改良して，文章推敲支援や文章作成支援に役立てていきたいと考えている．

## ◇ 参　考　文　献 ◇

[Barzilay 08] R. Barzilay and M. Lapata: Modeling Local Coherence: An Entity-Based Approach, Computational Linguistics, Vol. 34, No. 1, pp. 1-34 (2008)

[Burstein 04] J. Burstein, M. Chodorow, C. Leacock: Automated Essay Evaluation: the Criterion Online Writing Service, AI Magazine, Vol. 25 No. 3, pp. 27-36 (2004)

[藤井 08] 藤井敦: OpinionReader：意思決定支援を目的とした主観情報の集約・可視化システム，電子情報通信学会論文誌 D，Vol. J91-D，No. 2，pp. 459-470 (2008)

[Hotta 00] M. Hotta: Mapping Policy Discourse with CRANE: A Spatial Understanding Support System as a Medium for Community Conflict Resolution, Environment and Planning B: Planing and Design, Vol. 27, No. 6, pp. 801-814 (2000)

[石岡 08] 石岡恒憲: 日本語小論文の論理構成の把握とその図式表現, 人工知能学会論文誌, Vol. 23, No. 5, pp. 303-309 (2008)

[Knight 02] K. Knight, and D. Marcu: Summarization beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression, Artificial Intelligence, Vol. 139, No. 1, pp. 91-107 (2002)

[松本 02] 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)

[松村 03] 松村真宏, 加藤優, 大澤幸生, 石塚満: 議論構造の可視化による論点の発見と理解, 知能と情報, Vol. 15, No. 5, pp. 554-564 (2003)

[西田 06] 西田正吾, 伊藤京子, 仲谷美江: 市民のためのコミュニケーションを支援する情報システム, 電気学会論文誌 C, Vol. 126, No. 4, pp. 414-423 (2006)

[大澤 99] 大澤幸生, ネルス E. ベンソン, 谷内田正彦: KeyGraph：語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会論文誌, Vol. J82-D1, No. 2, pp. 391-400 (1999)

[大竹 02] 大竹清敬，岡本大吾，児玉充，増山繁: 自由作成要約に対応した新聞記事要約システム YELLOW, 情報処理学会論文誌「データベース」, Vol. 43, No. SIG2(TOD13), pp. 37-43 (2002).

[相良 07] 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)

[Smith 01] M. S. Smith and E. Vela: Environmental Context-Dependent Memory: A Review and Meta-Analysis. Psychonomic Bulletin & Review, Vol. 8, pp. 203-220 (2001)

[砂山 01] 砂山渡, 谷内田正彦, 文章の特徴を表すキーワードを発見して重要文を抽出する展望台システム, 電子情報通信学会論文誌, Vol. J84-D-I, No. 2, pp. 146-154 (2001)

[砂山 06] 砂山渡，橘啓八郎: サブトピックモデルに基づく文章の流れの評価指標の提案, 知能と情報, Vol. 18, No. 2, pp. 280-289 (2006)

[内田 97] 内田友幸, 田中英彦: 可読性向上を図る対話的文書自動彩色システム, 電子情報通信学会論文誌 D-II, Vol. J80-D-II, No. 12, pp. 3173-3180 (1997)

[魚崎 00] 魚崎祐子, 野嶋栄一郎: 下線ひき行為が文章理解に及ぼす影響, 日本教育工学雑誌, Vol. 24, pp. 165-170 (2000)

[Radev 02] D. R. Radev, E. Hovy, and K. McKeown: Introduction to the Special Issue on Summarization, Computational Linguistics, Vol. 28, No. 4, pp. 399-408 (2002)

〔担当委員：高間 康史〕

2009 年 1 月 26 日　受理

## 著　者　紹　介

**西原　陽子**（正会員）

2003 年大阪大学基礎工学部システム科学科卒業．2005 年同大学大学院博士前期課程修了．2007 年同大学院博士後期課程修了．博士（工学）．日本学術振興会特別研究員（DC1，PD）を経て，2008 年東京大学大学院工学系研究科助教，2009 年同講師，現在に至る．コミュニケーションデザイン，サービスシステムに関する研究に従事．電子情報通信学会，情報処理学会，各会員．

**佐藤　圭太**

2007 年広島市立大学情報科学部情報機械システム工学科卒業．2009 年広島市立大学大学院システム工学専攻博士前期課程修了．現在，キヤノン株式会社勤務．

**砂山　　渡**（正会員）

1995 年大阪大学基礎工学部制御工学科卒業．1997 年大阪大学大学院博士前期課程修了．1999 年大阪大学大学院博士後期課程中退．同年同大学院助手，2003 年広島市立大学助教授，2007 年同准教授，現在に至る．博士 (工学)．人間の創造活動を支援する研究，ならびにテキスト中の話題の流れ，発散と収束に関する研究に興味を持つ．言語処理学会，IEEE，各会員．