論 文

Technical Papers

修辞構造のアノテーションに基づく要約生成

Summarization of Multiple Documents with Rhetorical Annotation

綾 聡平
Sohei Aya
東京大学大学院情報理工学系研究科
School of Information Science and Technology, University of Tokyo
s-aya@miv.t.u-tokyo.ac.jp

松尾 豊
Yutaka Matsuo
産業技術総合研究所 情報技術研究部門 / 科学技術振興事業団 CREST
National Institude of Advanced Industrial Science and Technology / Japan Science And Technology Agency
y.matsuo@aist.go.jp, http://www.carc.aist.go.jp/~y.matsuo/

岡崎 直観
Naoaki Okazaki
東京大学大学院情報理工学研究科 / 産業技術総合研究所 情報技術研究部門
School of Information Science and Technology, University of Tokyo
okazaki@miv.t.u-tokyo.ac.jp

橋田 浩一
Kôiti Hasida
産業技術総合研究所 情報技術研究部門 / 科学技術振興事業団 CREST
National Institude of Advanced Industrial Science and Technology / Japan Science And Technology Agency
hasida.k@aist.go.jp, http://www.carc.aist.go.jp/~hasida/

石塚 満
Mitsuru Ishizuka
東京大学大学院情報理工学系研究科
School of Information Science and Technology, University of Tokyo
ishizuka@miv.t.u-tokyo.ac.jp, http://www.miv.t.u-tokyo.ac.jp/~ishizuka/

keywords: semantic authoring, rhetorical structure theory, automatic summarization, spreading activation

Summary

In this paper, we propose a new algorithm of summarization which targets a new kind of structured contents. The structured content, which is to be created by semantic authoring, consists of sentenses and rhetorical relation among sentences: It is represented by a graph, where a node is a sentence and an edge is a rhetorical relation. We simulate creating this content graph by using news paper articles that are annotated rhetorical relations by a GDA tagset. Our summarization method basically uses spreading activation over the content graph, followed by particular postprocesses to increase readability of the resultant summary. Experimental evaluation shows our method is at least equal to or better than Lead method for summarizing news paper articles.

1. はじめに

我々が生活していく上で，論文，レポート等は言うに及ばず，メモ書き，手紙，メール等，文章を用いて情報を他人に伝達する機会は数えきれない．しかしそれが上手くいかず，歯痒い思いをしている人は少なくないであろう．「文 A と文 B がこのような関係を持っているから，この順で並べるとよりわかりやすい」というような文の流れや構成を考えることは，憂鬱な作業のひとつであり，さらに，文を 1 次元化して並べることにより，伝えきれない情報があるはずである．

このような問題を解決するべく考えられているのが，**セマンティックオーサリング**である [Hasida 03]．これは，図 1 の例のように，単文として様々な情報を記述し，さらにその間にある関係をリンクで結ぶことでコンテンツを構造化し記述するものである．このグラフ形式のコンテンツは順序なしであるので，線形に並べるコストがなくなり，書き手が書きやすいという利点がある．このような形式で構造化することで，検索や要約，発想支援などに有用であることが期待されている．実際，構造化により文書検索の精度があがるという結果が得られている [Miyata 02]．もちろん，単文の持つ意味は文脈に依存するなどの難しい問題はあるが，できる限り文の関係性を明示化することによって，文書処理を実現可能な手段で意味的に深めようというのが我々のアプローチである．

セマンティックオーサリングは，文書に代わる新しい知識の記述手段として，大人数による効率的な知識共有や蓄積のために利用されることが想定されている．その際，多くの人がさまざまな内容についてコンテンツを記述することになるので，膨大なグラフからユーザの必要とする情報をコンパクトに取り出す技術が必要になる．これは，グラフで表されるコンテンツからの要約と言いかえることもできる．

一方，自動要約の分野では，近年，複数文書要約の研究

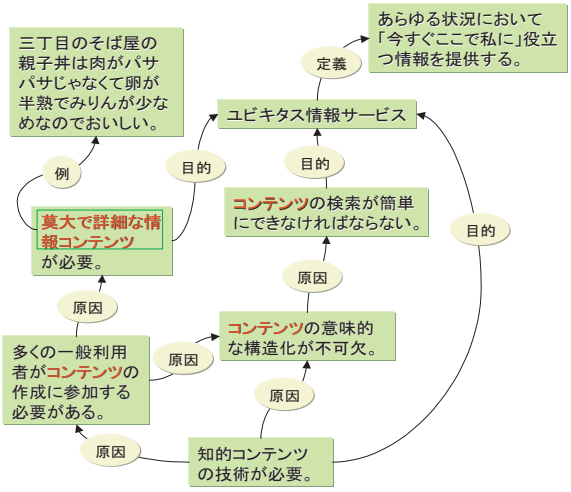


図 1 セマンティックオーサリングによるコンテンツの例

```
<?xml version="1.0" encoding="UTF-8"?>
<gda lang="jpn">
  <su id="node01" addition="node02">
    <orgnamep id="nihonTV">日本テレビ放送網（本社･東京都港区）</orgnamep>の<n id="producer">男性社員プロデューサー（４１）</n>が，<orgnamep id="videoresearch">視聴率調査会社「ビデオリサーチ」（本社・東京都中央区）</orgnamep>の調査対象世帯に現金などを渡して自分が制作した番組を見るように依頼していた　と<datep value="2003-10-24">２　４　日</datep>，<orgname eq="nihonTV">日テレ</orgname>が発表した．
  </su>
  <su id="node02">
    調査対象世帯は外部には秘密になっているが，<np eq="producer">プロデューサー</np>は，興信所を使って割り出していた．
  </su>
</gda>
```

図 2 GDA によるアノテーション例

が活発に行われている．いくつかの手法では，文や語のネットワークをもとに要約を生成しており [Hasida 87, Mani 97, Okazaki 03]，セマンティックオーサリングで得られるグラフ*1と関連が深い．基本的には同様の手法が適用可能であると考えられる．

本研究では，以上のような背景に鑑み，次のようなアプローチを取る．まず，あるトピックに関する複数文書の新聞記事を集める．この中で，文間の関連をもとに，図 1 のような文をノードとするグラフを構成する．その上で，グラフを用いた要約でよく用いられる活性拡散のアルゴリズムを用い，要約を生成する．

通常の要約の研究では，自然言語で書かれた文書を対象としており，文間の関係が明示的に記述されている構造化された文書を対象とはしていない．本研究ではセマンティックオーサリングで得られたグラフを想定し，文間の修辞関係まで明示的に記述されている構造化文書を仮定している．仮定が異なるため，単純な比較はできないが，本稿ではその全体の処理の概要を述べるとともに，比較実験を行う．

本稿は全 6 章で構成されている．2 章では，関連研究として，GDA および修辞構造理論について簡単に説明する．3 章で提案手法の詳細について述べ，その評価を 4 章に示す．5 章で本手法に関連する話題について議論を行う．

2. 関 連 研 究

2･1 大域文書修飾：GDA

本研究では，**GDA**（大域文書修飾：Global Document Annotation）タグ集合 [橋田 98, 橋田 02] を利用する．GDA を用いることで，文書中の統語照応構造，修辞構造，対話構造，語義等の情報を明示的に示すことができる．意味構造を明示するコンテンツ記述のための枠組みとしては，他に UNL[Uchida 00] や OWL*2などがあるが，GDA タグ集合の特徴は原コンテンツ（文書に限らずビデオやオーディオや実世界など，人間が理解できるコンテンツ）と意味構造を結びつける点にある．

GDA で記述した文書の例を図 2 に示す．GDA のアノテーションには様々な粒度が認められている．この例は比較的粗い粒度のアノテーションである．

2･2 修辞構造理論を利用した要約作成

本研究で対象とする文間の関係は，言語学では修辞構造理論（Rhetorical Structure Theory: 以下 RST）[Mann 87, Mann 88] と呼ばれるものに近い．ここではまず，RST について説明する．

RST は，文（あるいは節）間の関係に対するテキスト一貫性のひとつの理論であり，400 以上の短いテキストを分析した結果として開発された．大きな特徴は，テキスト中の修辞関係の大半が非対称であり，一方のテキストセグメントは他方に対して補助的であると考える点にある．文章の主旨に対してより中心的役割を果たすテキストセグメントを核，補助的な役割を果たすセグメントを衛星と呼ぶ．文書全体の修辞構造は，RST 木とよばれる木構造を成すことになる．

例えば，次の文があったとする．

> Ask for SYNCOM diskettes, with burnished Ectype coating and dust-absorbing jacket liners[2]. As your floppy drive writes or reads,[3] a Syncom diskette is working four ways to keep loose particles and dust from causing soft errors, dropouts.[4]

*1 通常，エッジに距離や強さの値が付与されたものをネットワーク，そうでないものをグラフと言い，本研究で扱うのはグラフだが，本論文中では，読みやすいようにネットワークと表現する場合もある．

*2 http://www.w3.org/TR/owl-features/

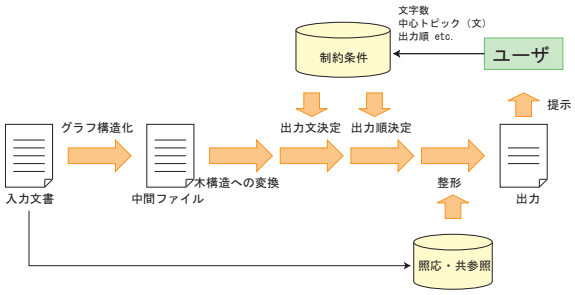


図 3 システム概観

文 (3) と文 (4) は RST では「環境」の関係にあり，文 (3) が衛星，文 (4) が核である．つまり，文 (3) が文 (4) の環境を示している．さらに，文 (2) と文 (3)(4) は「動機」の関係にあり，文 (2) が核，文 (3)(4) が衛星である．文 (3)(4) が文 (2) の動機を表している．したがって，文 (2)(3)(4) 全体では，最も中心的な役割を果たすのが文 (2) であり，1 文だけ選ぶ場合には，文 (2) が要約としてふさわしいことになる．

RST に基づく要約の手法 [Marcu 99b] では，各々のテキストセグメント間に修辞関係があると仮定し，その関係を自動的に取得する．一旦，木構造が得られれば，あとは木構造にしたがって決定されるテキストセグメントの半順序関係にしたがって，重要なものから出力していけばよいことになる．この手法では人間が構築した修辞構造木に対し，78%の精度，67%の再現率で重要文の抽出ができると述べられている．

3. 修辞構造に基づく複数の構造化文書要約

本章では，文間の関係を明示的に記述した文書から，要約を生成する手法について述べる．

3·1 本研究における修辞関係

本研究は RST を参考にしているが，いくつかの点で拡張を行っている．

まず，本研究で対象とする修辞関係は，文書内で隣り合った文や近接する段落に記述されているものとは限らない．例えば,「動機」や「例」の関係にある 2 つの文が，文書内で遠く離れていることもあれば，このような関係が複雑に交差していることもあり得る．したがって，本研究では修辞関係の分類を参考にするものの，その関係は 2 つの文間にさまざまに存在すると考える[*3]．

さらに，本研究では複数文書を扱うが，異なる文書における文間にも修辞関係が存在すると仮定する[*4]．これは，通常の複数文書要約では想定できないが，本研究では，セマンティックオーサリングによって得られる構造化されたコンテンツの代替として，複数文書の文が織りなすグラフ構造を捉えているため，このような仮定を置く．また，修辞構造はテキスト一貫性に由来するものであり，複数文書全体がひとつの RST 木になるということは考えにくいため，本研究では，修辞関係の階層は，多くても 2 から 3 段階にとどめている．

本研究では，これまでの修辞構造理論を参考に，35 セットの修辞関係を定義した．修辞関係は，逆を考慮しないもの（背景等），対称な関係（換言等），非対称な関係（原因と結果等）の大きく 3 つに分けることができる [綾 04]．その一部を下に示す．

逆関係を考えない修辞関係 addition, background, circumstance, manner

対称な修辞関係 antithesis, comparison, contrast, disjunction, dissimilar, list, proportion, restatement, sametime, similar

非対称な修辞関係．括弧内は逆関係 after(before), attribution(content), cause(result), comment(topic), conclusion(evidence), condition(outcome), sequence(invertedsequence) ほか

将来的には修辞関係の細かい区別をアルゴリズムに反映していく必要があるが，ここでは，文の出力順の決定および整形処理以外には区別せず，関係があるかないかだけに着目する．

まず，入力文書からアノテーションで与えられた文間の修辞関係を抜き出し，文をノードとするグラフを作成する．次に，このグラフ構造を用いて活性拡散を用いた計算を行い，ノードの重要度を評価した上で要約の中心となる文を決定する．中心となる文と修辞関係にある文から出力する文を選び，その順序を定め，整形した上で最終的な出力とする．本システムのおおまかな流れを図 3 に示す．

3·2 入 力 文 書

本研究では，新聞記事を入力文書としている．要約の研究対象としては一般に新聞記事が用いられることが多く，また，[Marcu 99a] によると，人間が修辞関係を判断した際に判定者間の一致度が高いという特徴がある．このように安定した修辞関係が得られる複数の記事を用いることで，セマンティックオーサリングで構築される構造化コンテンツにできるだけ近いもの[*5]が得られると想定している．

GDA によってタグ付けされた複数の文書から，まず修辞構造をもとにひとつのグラフを生成する．すべての文をノードに，明示的に記述されたすべての修辞関係をリンクとする[*6]．

*3 RST では，修辞関係は通常は隣接するが，例外的に隣接しない場合もあり得るとされている．

*4 厳密には，もはやテキスト一貫性の根拠ではないので，修辞関係とは言い難い．

*5 新聞記事の各文は前後の文との整合性や読みやすさといった点から整形されているが，セマンティックオーサリングで構築される構造化コンテンツにおいては，整合性や読みやすさよりも，内容を的確に表すところに主眼が置かれている．

*6 なお，段落のような複数の文のかたまりをひとつのノードと

```
<?xml version='1.0' encoding='UTF-8'?>
<file>
  <Nodes>
    <Node nodeID='node01'>
      <su>日本テレビ放送網の男性社員プロデューサーが
視聴率調査会社ビデオリサーチの調査対象世帯に現金などを
渡して自分が制作した番組を見るように依頼していたと２４
日，日テレが発表した．</su>
    </Node>
    <Node nodeID='node02'>
      <su>...</su>
    </Node>
  </Nodes>
  <Relations>
    <Relation source='node01' target='node02'
                rst='result' inv_rst='cause'/>
  </Relations>
</file>
```

図 4 複数文書をまとめた GDA 中間ファイル

セマンティックオーサリングにより出力される形式との整合性をとるため，このグラフ構造を XML の中間ファイルとしていったん生成する（図 4）．この図では，簡単のため，文を示す su 属性以下のアノテーションは全て省略している．Node 要素の子ノードとして，ノードの持つテキストを表現し，Relations 要素内 Relation 要素でノード間にあるリンクを記述する．source 属性はリンク元，target 属性はリンク先である．rst 属性は source から target へのリンクを表現するときの修辞関係であり，inv_rst 属性は target から source へのリンクを表現するときの修辞関係を示している．

3·3 活性拡散によるノードの重要性計算

本手法では，文間の関係を表すグラフに活性拡散を適用し，顕現性の高い文を抽出することにより重要文の抽出を行う．

[Skorochod'ko 72] では，テキスト結束性に着目し，2 つの文中に同一語や類義語，上位語を含むかなどによって，文間のグラフ構造を生成し，その解析を行っている．また，[Mani 97] では，照応や反復表現から語のグラフ構造を作成し，その上で活性拡散により重要な語を含む文を取り出している．

活性拡散による重要文抽出は，一般にはテキスト結束性に着目した手法であるが，テキスト一貫性に関する理論である RST に着目した本手法が，活性拡散による重要文抽出を行うのは次のような理由のためである．

- セマンティックオーサリングによって生成される文間のグラフ構造はあくまでも局所的なものであり，全体が木構造になるわけではないため，RST 木による要約法は利用できない．
- 段落や文，句，語の関係性を表すネットワークが与えられたときに，ノードの重要度を計算する方法として，エッジの多さや活性拡散による活性値を用いる方法が一般的である [奥村 99, Mani 03]．
- 活性拡散を使った要約では，活性値の高いノードの周りのノードも活性値が高くなることが多く，中心的なトピックを重点的に取り出す傾向が強い．本手法では，修辞関係でエッジを張っており，近接したノードの方が出力しやすいため，この性質は望ましい．

活性拡散（Spreading Activation Model）は，ネットワークにおけるノードの重要性を計算する方法である [M.R.Quillian 68, J.R.Anderson 83]．ノード数を n，反復を t とすると，各ノードのもつ活性値が収束するまで以下の計算を繰り返す [Huberman 87]．

$$\mathbf{A}(t) = \{(1-\rho)\mathbf{I} + \rho\mathbf{R}\}\,\mathbf{A}(t-1)$$

ここで $\mathbf{A}(t)$ は反復 t における活性値を表す n 次元のベクトル，$\mathbf{I}$ は $\mathbf{A}(t-1)$ の活性値を $\mathbf{A}(t)$ に伝搬させる単位行列，$\mathbf{R}$ はネットワークの構造を表す $n \times n$ の伝搬行列であり，$\mathbf{R}$ の i 行 j 列の要素 R_{ij} はノード i とノード j の関連の強さを表す．本手法では修辞関係の違いによる影響の大小を考慮しないため，0-1 の非対称行列である．(対角成分は 0)．また，ρ は隣接するノードからの影響の強さを調整するパラメータである．$\mathbf{A}(0)$ はすべての要素を 1 とする．

最終的に，収束した時点でのベクトル $\mathbf{A}(t)$ が各ノードの活性値を表すが，本手法では，活性値をそのままノードの重要度として用いるのではなく，活性値と修辞構造を考慮してノードの重要度を計算する．この処理は，試行錯誤を行いながら次のように定めた．

まず，自分自身のノードの活性値だけでなく，近接ノードの活性値も利用する．これは，活性値の低い多くのノードと関連があることによって活性値が高くなっているノードを取り出すのではなく，周りのノードも活性値が高く自分自身も活性値が高いノードを取り出すためである．つまり，グラフにおいて，よりトピックの中心的なノードを選択していることに相当する．さらに，短いテキストで多くの修辞関係があるために活性値が高くなっているノードより，長いテキストを持つノードの方が中心的なノードとしてはふさわしいため，テキスト長も重要度の式に織り込んだ．

ノード n の重要度 w_n は，以下の式で計算する．

$$w_n = \left(a_n + \sum_{k \in K_n} \frac{a_k}{d_{nk}}\right) \times \sqrt{l_n}$$

なお，a_n はノード n の活性値，K_n は n からパス長 3 までのノードの集合，d_{nk} はノード n からノード k までのパス長，l_n はノード n の持つテキストの長さである．

して扱う処理，段落と文の間にリンクを生成する処理も行っている [綾 04]．

node ID	記事 ID-位置	出力文
31	1-2	日本テレビ放送網（本社・東京都港区）の男性社員プロデューサー（４１）が，視聴率調査会社「ビデオリサーチ」（本社・東京都中央区）の調査対象世帯に現金などを渡して自分が制作した番組を見るように依頼していた．
06	1-9	日テレは，平均視聴率の「４冠王」を９年連続で獲得している．
28	1-47	プロデューサーは 1984 年入社で，1991 年にスポーツ局から編成局の制作部門に自ら希望して異動し，バラエティー番組を中心に手がけていた．
22	1-14	プロデューサーの買収工作は 2002 年 07 月から行われた．
08	1-15	調査対象世帯の割り出しを，埼玉県内の興信所に依頼した．
13	1-24	2003 年 01 月，同年 04 月，同年 09 月のプロデュース番組でも「買収工作」をしており，それぞれ４世帯ほどが応じた．
10	1-19	割り出した世帯には，知り合いの元番組制作会社の社長夫婦を介して，2002 年 09 月 19 日放送の「芸能人犯罪被害スペシャル」と同月 26 日の「奇跡の生還　芸能人版」を見るよう依頼．
11	1-20	承諾した４世帯に，５千円から１万円の商品券か現金を渡した．
37	2-17	プロデューサーの買収工作を仲介していたとされる元番組制作会社社長は，2003 年 02 月，ビデオリサーチ社側から強い抗議を受けて仲介役をやめていた．

表 1　日本テレビプロデューサによる視聴率買収問題の要約例

3·4　出力文と出力順の決定

要約の中心となるノードは，重要度 w_n の最も高いノード n である．この中心ノードから修辞関係のエッジを張られたノードを候補とし[*7]，その中から最も活性値の高いものを出力対象のノードに加えていく．これを再帰的に字数制限に達するまで繰り返す．したがって，出力対象となる全てのノードは，中心ノードから何らかの修辞関係をたどっていけば到達できることになる．

次に，出力対象のノードに対し，どちらを先に出力するかという順序を決定する．ある node1 と node2 にエッジがある場合，node1 を node2 の前方に出力するか，後方に出力するかを，修辞関係の種類によってあらかじめ決めておいた順序により定める．例えば，node1 が node2 と sequence の関係にあれば node1 を前，node2 を後ろにする．逆に invertedsequence の関係にあれば，node1 を後ろ，node1 を前にする．他にも，example の関係にあるものは後ろ，condition の関係にあるものは前など，修辞関係のいくつかについては順序を定めておく．それで判断できない場合には，文の属する文書の日付けによって順序を決める．

状況によっては，文の出力順序が完全に決まらない場合がある．例えば，node1 から node2 と node3 にエッジが張られており，両方が node1 より前もしくは後に出力することは分かっているが，node2 と node3 には前後関係の判断ができない場合である．この場合，修辞関係が少ないノードを優先する．つまり，複雑な構造を持つ文の集合を後から述べることになる．

3·5　整　　形

以上のようにして出力順序まで決定した後，より読みやすい要約にするために文の整形を行う．

まず，文字数を減らすために冗長な主語の省略を行う．GDA によって示された主語が，先の文の主語と同じである場合には，その主語とそれに係る修飾語を全て削除する[*8]．

次に，GDA で照応・共参照が明示的に示されるので，これを利用した整形を行う．GDA では，例えば「日本テレビ放送網」「日テレ」「同社」等に同じ id を割り振ることで同じ対象を指すことを表す．したがって，同じ ID を持つ語に対し，初めてその語が現れる場合には正式な表現を，2 回目以降に現れる場合には通称を使用する．なお，正式な表現とは最も長い表現，通称とは原文中で最も使用回数の多い表現と定義する．例えば,「日本テレビ放送網」が最も長い表現なので正式名称,「日テレ」が最も多く出現するので通称である．さらに，他の対象を指す照応・共参照の表現を挟んでいない場合は,「同＊＊」（同社，同市など）という表現に変換する．

GDA のアノテーションにより，日付に関しても表現が指し示す正確な日付が取得できる．したがって，対象とする文の日付が，それまでに述べられている文の日付と異なるならば，相違部分以降の正確な年月日を出力する．例えば,「2004 年 1 月 26 日」という表現があり，次に 2003 年 12 月 24 日を示す表現が来れば「2003 年 12 月 24 日」と変換し，また 2004 年 1 月 25 日を示す日付が来れば「同月 25 日」と変換して出力する．

---

*7　修辞関係の種類により，たどるエッジについても制限を設けている．例えば,「換言」は冗長なのでたどらない．

*8　主語を表すタグがついていない場合には，形態素解析器（茶筌 [松本 00]）と構文解析器（南瓜 [工藤 02]）により主語を特定する．例えば，が格，は格の格助詞に続く名詞を主語とするなどの簡単な処理も可能である．

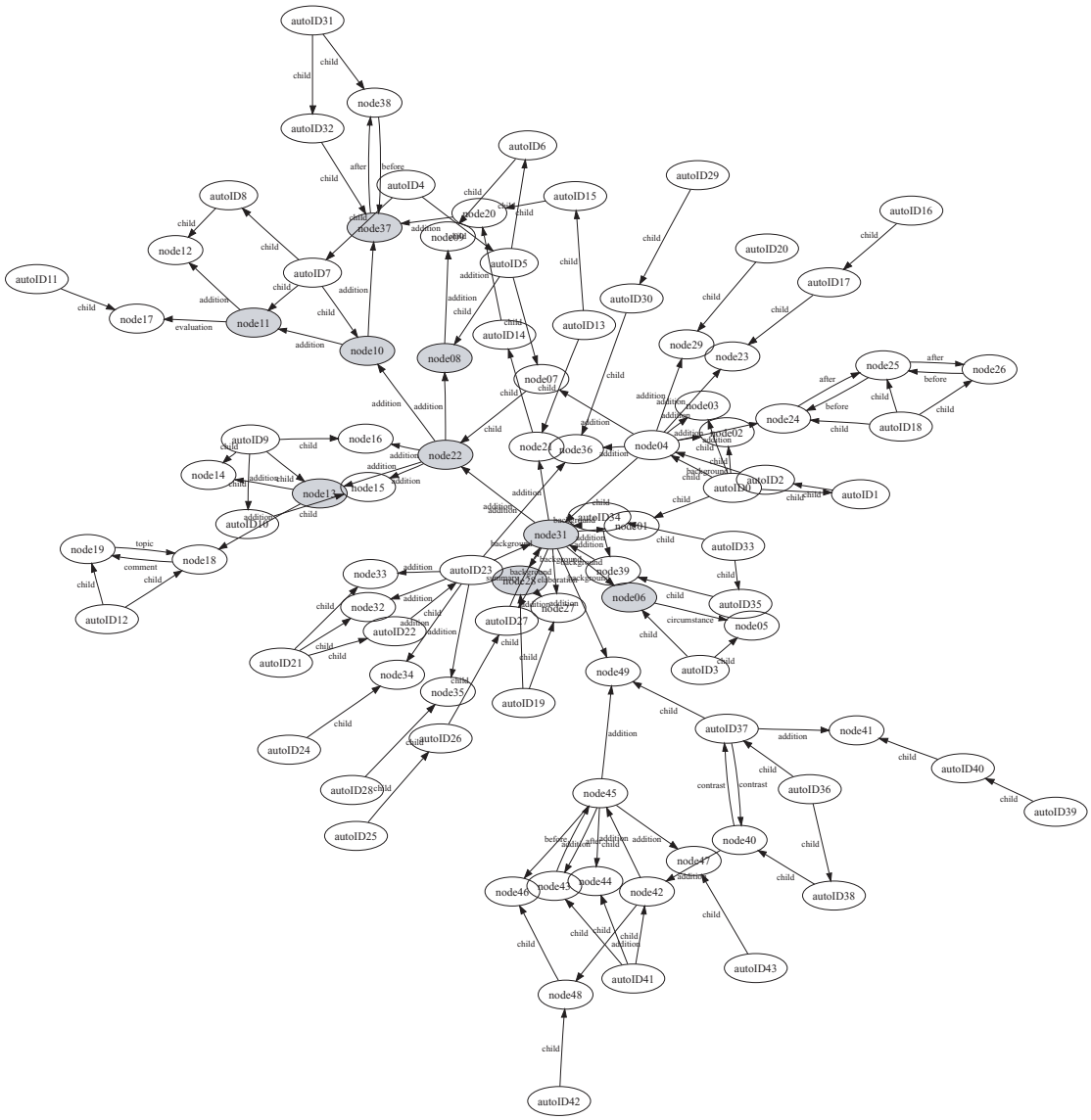


図 5 原文の修辞構造

4. 結　　果

4·1 出　力　例

要約の例を表 1 に示す．アサヒ・コム*9 の日本テレビプロデューサによる視聴率買収問題に関する 3 記事*10に，GDA のアノテーションを施し，要約したものである．およそ 20%の要約率である．この記事のもとの修辞構造を図 5 に，要約したものを図 6 に示す．もとの修辞構造の中で中心的なノード (node31) が要約の最初のノードとして選ばれている．

原文は，修辞構造，照応・共参照，日付については，完全にアノテーションを施している．比較手法と条件を合わせるため，それ以外の格構造や依存構造についてはアノテーションを利用しない．表の一番左の項目は nodeID を表し，2 番目の欄は記事 ID および原文中で何文目に出現するかを示す．原文のトピックをカバーしながら，修辞関係において関連ある文が隣り合っており，読みやすい要約が得られている．また，原文で隣り合った文が比較的隣り合って出力されやすいことも読み取れる．

4·2 評 価 実 験

本手法の有効性を調べるため，評価実験を行った．実験対象の文書集合は，2001 年から 2002 年の NTCIR Workshop3*11 における Automatic Text Summarization Task 2（TSC2）のデータの一部を使用した．1998 年から 1999 年の毎日新聞の新聞記事である．用いたデータは 1459 文字，2214 文字，2781 文字の 3 記事である．

次のシステムで比較を行った．

システム 1　人間による要約
システム 2　LEAD 法＋重要文抽出による要約
システム 3　提案手法（照応・共参照，省略の処理なし）

*9　http://www.asahi.com/
*10　それぞれ 2003 年 10 月 25 日 0 時（28 文），10 月 25 日 17 時（11 文），10 月 28 日 7 時（7 文）である．全文は紙幅の都合で載せることができないが，[綾 04] を参照．
*11　http://research.nii.ac.jp/ntcir-ws3/

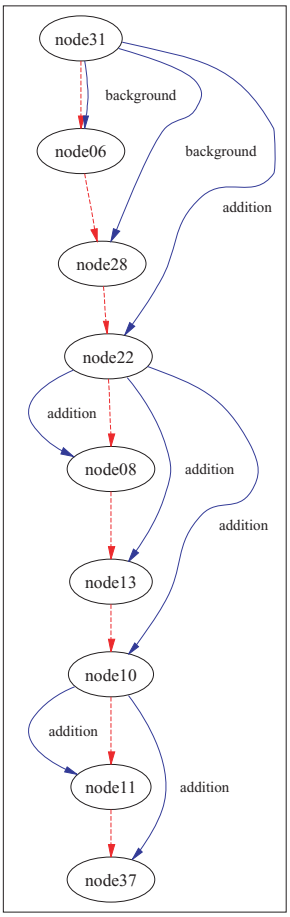


図 6 要約結果の修辞構造

システム 4 提案手法（照応・共参照，省略の処理あり）人間による要約は，TSC2 のものを，そのまま使用した．

LEAD 法は，文書の初めの数文を要約として取り出すもので，最もポピュラーな要約手法のひとつである．特に新聞記事において良い成績を上げることが知られているが，本研究では複数文書を対象としているため，どの記事から取ってくるかという自由度が残る．そこで，TF·IDF[Harman 92] を用い，各語の TF·IDF 値を足し合わせることで文の重要度とした．全ての記事の先頭の文を候補とし，重要度の最も高い文をピックアップするという処理を，字数制限に達するまで繰り返す．文の並び順に関しては，原文の通りの順序および日付を利用した．

提案手法は，GDA による照応・共参照や主語の省略などの整形処理を行わないシステム 3 と，処理を行うシステム 4 の両方を用いた．

評価方法は TSC2 にならい，各手法で長い要約（字数 500 文字以内）と短い要約（字数 250 字以内）の 2 種類を生成し，それぞれの要約に対し，内容と読みやすさについて評価してもらう方法をとった．被験者は 11 名[*12]で，各項目に 4 段階で答えてもらった．

*12 情報系の大学生および大学院生．

4·3 結 果 と 考 察

評価結果を図 7〜図 10 に載せる．各評価者の平均を評価値としている．

まず，システム 3 とシステム 4 に関しては，全ての項目で，照応・共参照と主語等の省略の処理を行っているシステム 4 が，行わないシステム 3 よりも評価が高かった．照応・共参照や省略の処理を行うことで結束性が上がり，文のつながりが良くなって読みやすくなると考えられる．この効果は，短い要約よりも長い要約の場合に顕著である．

また，システム 3 とシステム 4 は基本的には同じ文を抽出するが，システム 4 では主語等の省略を行っているため，字数制限までにもう 1 文入る可能性がある．そのために，システム 4 は内容点でシステム 3 よりも評価値が高くなっている．しかし，基本的には同じ文を出力しているので，可読性が評価者の内容に関する評価にまで影響を与えている可能性も考えられる．

システム 3 と LEAD 法（システム 2）を比較した場合，短い要約では内容点，可読性ともにほとんど差が無いものの，長い要約では両者とも大幅に評価値が低い．LEAD 法は，簡単ではあるが新聞記事に有効な方法であり，単に修辞関係を考慮しただけでは LEAD 法を上回る結果にはなっていない．

一方，システム 4 と LEAD 法を比べると，同じ，もしくはシステム 4 の方がやや高い評価値を得ている．文の間にある修辞関係が理解しやすくなり，評価が高まったと予想できる．この評価からは，本手法が LEAD 法に比べて十分な優位さがあるとは言い難いが，修辞構造に着目し，新聞記事に特有の処理を最小限に抑えたシステムで LEAD 法と同等の結果になったことは，少なくとも，提案システムがある程度の実用性を持つ要約システムであることを示しているだろう．文間の修辞関係が与えられた複数文書，ないしはセマンティックオーサリングによって生成された構造化コンテンツという新しいデータを用いて，ひとつの要約システムが構築できたことを意味すると考えている．

本システムでは，評価対象として新聞記事を用いたが，セマンティックオーサリングで記述されるコンテンツは本来はさまざまな内容に渡るものであり，新聞記事に対する評価が一般的なコンテンツに対する要約性能を示しているとは限らない．しかし，要約手法の評価としては，現在のところ新聞記事を用いるのが標準的であること，また構造化コンテンツに対する適切な比較手法がないことから，本論文では新聞記事を対象とし，LEAD 法，人間による要約との比較を示している．今後，セマンティックオーサリングの研究が進むにつれて，さまざまなコンテンツに対する要約を比較対象としていく必要があると考えている．

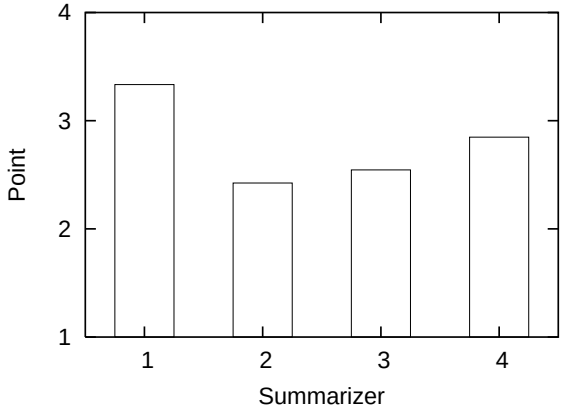


図 7 短い要約の内容に関する評価

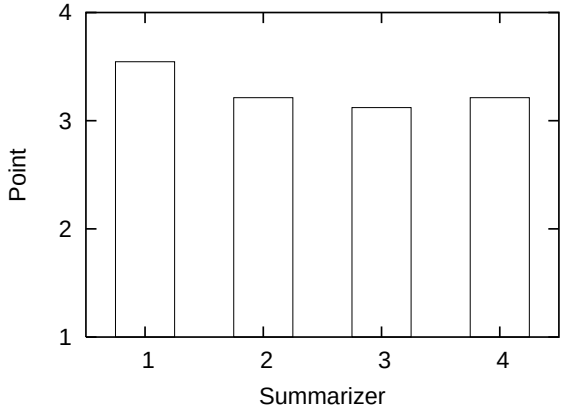


図 8 短い要約の可読性に関する評価

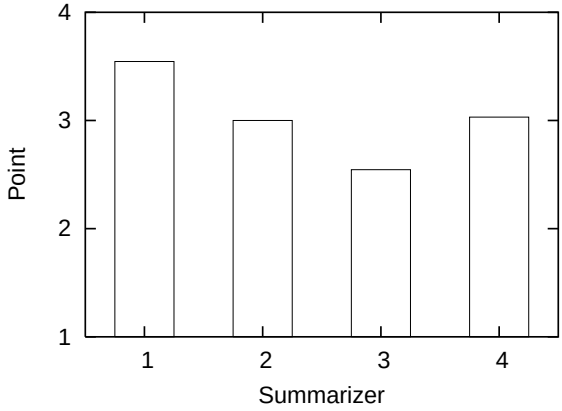


図 9 長い要約の内容に関する評価

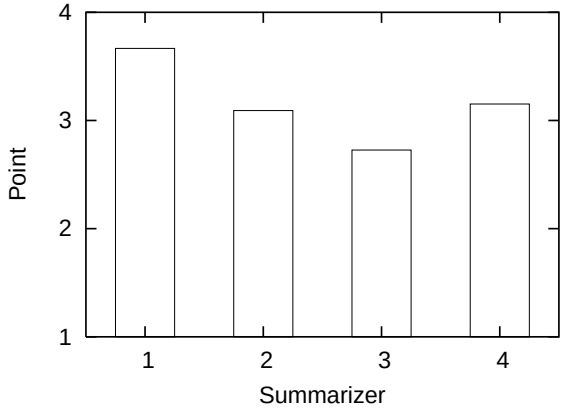


図 10 長い要約の可読性に関する評価

5. 議　　論

5·1 ユーザとのインタラクション

本システムは，セマンティックオーサリングと合わせて使うことが想定されており，ユーザとのインタラクションを考慮した要約システムとなっている．具体的には，ユーザが着目するノード（文）を中心に関連した情報の要約ができる．これは，活性拡散における活性値の初期値をユーザが指定したノードに与えることで可能である．また，指定したノードからの修辞関係を考慮して文の整形処理を行う．

例として，表 1 と同じ記事集合および要約率に対して，ユーザがひとつの文を指定した場合に出力される要約を図 11 に示す.（1 文目が指定した文である.）この文は，日本テレビ社が買収工作について記者会見で発表したという内容であり，得られる要約も日本テレビ社による発表に関わる情報が中心になっている．このように，同じ記事集合でも，ユーザの指定したノードにあわせて，異なる要約を出力することができる．

5·2 アノテーションのコストと効果

本システム（システム 4）では，新聞記事に特有の処理を最小限にしながら LEAD 法と同じかやや上回る評価を獲得している．しかし，アノテーションをつけるコスト，すなわち時間と労力まで考えると，LEAD 法と比較しての差は見合わないという考え方もあるだろう．また，人間による要約と比較した場合には，依然として大きな差がある．

しかし，アノテーションは，要約精度を上げるためだけに行うのではない．本研究は，セマンティックオーサリングにより作成されたコンテンツから，ユーザの必要とする文章をある程度自動的に生成することを最終的な目標としている．その点からは，LEAD 法と同等以上の要約の性能というのは，ひとつの側面からの評価にすぎない．他にも，GDA 文書の検索に関しては，従来文書より高効率である [Miyata 02] という効果も示されており，翻訳，質問応答，知識発見などの応用 [橋田 02] についても研究が進められている．

本研究で用いる新聞記事は，人手で修辞関係に関するタグが与えられている．[Marcu 99a] で述べられているように，手法自体の問題と実装上の問題を明確に分けるために，まずは人手によるタグの施された文書を用い，手法自体の可能性を探ることを目的としている．この処理を自動的に行う，またツールを用いて記述しやすい環境を構築するなど，タグ付けを支援するしくみについては本論文の範囲外であり，ここでは扱わない．

日本テレビ放送網（本社・東京都港区）の男性社員プロデューサー（４１）が，視聴率調査会社「ビデオリサーチ」（本社・東京都中央区）の調査対象世帯に現金などを渡して自分が制作した番組を見るように依頼していたと2003年10月24日，日テレが発表した．同社の萩原敏雄社長は記者会見を開いて陳謝したが，組織的な関与は否定した．同社では，調査委員会を設置し，事実関係を明らかにした上で社員の処分などを決める．萩原社長によると，プロデューサーの買収工作は2002年07月から行われた．ビデオリサーチ社は同年末までに，３世帯に対して特定の番組視聴などの働きかけがあったことを知り，すでに調査対象から外しているが，今回の日テレの発表を受け，改めて関東地区の全６００世帯について１，２カ月のうちに調べる．2003年10月24日の会見で萩原社長が組織ぐるみの関与を否定する一方で，他のプロデューサーも「調べるだけは調べる」としており，合わせて調査を進める．この問題は同月23日夜，一部のマスコミから同社に事実関係の確認取材があったことで，明らかになった．同社幹部がプロデューサーを呼び事情聴取.

図 11 中心ノードを変えた要約結果

5·3 システムについて

本研究で構築した要約システムは，セマンティックオーサリングを取り巻く一連のツールのひとつとしての位置づけであり，本システムの構築に伴って以下のツール群も開発した．

GDA アノテーション補完ツール GDA のアノテーションがある程度施されている文書に対し，茶筅と南瓜を用いて形態素解析，係り受け解析の結果をつけ加えることができる．

要約結果を表示するユーザインタフェース ユーザが中心としたい文を指定する，得られた要約をネットワーク図の形で見るなどの機能を持つ．

本研究では，セマンティックオーサリングにより作成されたコンテンツへ利用することを念頭に，新聞記事特有の処理は最小限にとどめている．また，複数の新聞記事の修辞関係を集約した中間ファイルを介すことで，セマンティックオーサリングにより作成されたコンテンツと同じ形式を扱えるように配慮している．

6. お わ り に

本論文では，セマンティックオーサリングにより作成されたコンテンツからの文章生成という最終目標への第一段階として，修辞関係等を明示的に与えた複数文書に対し，要約を作成する手法を提案した．新聞記事に対して，LEAD 法と同等以上の評価が得られることを示した．

本研究は，ユーザが指定したノードを中心とした要約を表示するなど，セマンティックオーサリングを取り巻くツールのひとつという位置づけである．今後の課題として，セマンティックオーサリングにより得られたコンテンツへの適用が挙げられ，現在，オーサリングツールの開発が進行中である．

また，セマンティックオーサリングでは，さまざまなトピックについてのコンテンツが想定されるため，新聞記事以外への本手法の適用と評価が必要であろう．さらに，本手法のひとつのメリットである，ユーザとのインタラクションによる要約の評価を行う必要がある．しかし，いずれも，現在の要約研究では一般的な評価方法が確立していないため，今後，評価方法に関する進展を見ながら検討していく予定である．

セマンティックオーサリングは，ユーザがその良さを実感し，実際にアノテーションされた文書が増えていく環境を実現するまで，理論やアルゴリズムの研究とともに，さまざまな使いやすいツールの開発や改良が必要である．大きなプロジェクトではあるが，本研究はその道標のひとつとなると考えている．

◇ 参 考 文 献 ◇

[綾 04] 綾 聡平：アノテーション付き多文書データからの要約生成, Master's thesis, 東京大学 (2004)

[Harman 92] Harman, D.: Ranking Algorithms, in *Information Retrieval: Data Structures and Algorithms*, pp. 363–392, Upper Saddle River, New Jersey: Prentice-Hall (1992)

[Hasida 87] Hasida, K., Ishizaki, S., and Isahara, H.: A Connectionist Approach to the Generation of Abstracts, in Kempen, G. ed., *Natural Language Generation: New Results in Artificial Intelligence, Psychology, and Linguistics*, pp. 149–156, Martinus Nijhoff (1987)

[橋田 98] 橋田 浩一：GDA — 意味的修飾に基づく多用途の知的コンテンツ —, 人工知能学会誌, Vol. 13, No. 4, pp. 528–535 (1998)

[橋田 02] 橋田 浩一：インテリジェントコンテンツ, 情報処理, Vol. 43, No. 7, pp. 780–784 (2002)

[Hasida 03] Hasida, K.: Distributed Semantic Authoring as Foundation of Semantic Society, in *Notes on From Semantic Web to Semantic World workhop conjoint with JSAI2003* (2003)

[Huberman 87] Huberman, B. A. and Hogg, T.: Phase Transitions in Artificialn Intelligence Systems, *Artificial Intelligence*, Vol. 33, No. 2, pp. 155–171 (1987)

[J.R.Anderson 83] J.R.Anderson, : A Spreading activation theory of memory, *Journal of Verbal Learning and Verbal Behavior*, pp. 261–295 (1983)

[工藤 02] 工藤 拓, 松本 裕治：チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842 (2002)

[Mani 97] Mani, I. and Bloedorn, E.: Multi-document Summarization by graph search and matching, in *Proc. of AAAI-97*, pp. 622–628 (1997)

[Mani 03] Mani, I.: 自動要約, 共立出版 (2003), 奥村 学, 難波 英嗣, 植田 禎子 訳

[Mann 87] Mann, W. and Thompson, S.: Rhetorical Structure Theory: A Framework for the Analysis of Texts, Technical report, Technical Report ISI/RS-87-185, Marina del Rey, California (1987)

[Mann 88] Mann, W. and Thompson, S.: *Rhetorical Structure Theory: Towerds a Functional Theory of Text Organization*, chapter 8, pp. 243–281, Text (1988)

[Marcu 99a] Marcu, D.: The automatic construction of

large-scale corpora for summarization research, in *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 137–144, New York: Association for Computing Machinery (1999)

[Marcu 99b] Marcu, D.: Discourse trees are good indicators of importance in text, in Mani, I. and Maybury, M. eds., *Advances in Automatic Text Summarization*, pp. 123–136, Cambridge, Massachusetts (1999), MIT Press

[松本 00] 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：日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書, Technical report, a (2000)

[Miyata 02] Miyata, T.: Information retrieval system based on graph matching, in *ECAI2002 workshop on Ontology Knowledge Transformation for the Semantic Web*, p. 109 (2002)

[M.R.Quillian 68] M.R.Quillian, : Semantic Memory, *Semantic information processing*, pp. 227–270, MIT Press, (1968)

[Okazaki 03] Okazaki, N., Matsuo, Y., Matsumura, N., and Ishizuka, M.: Sentence Extraction by Spreading Activation through Sentence Similarity, *IEICE Transactions of Information and Systems*, Vol. E86-D, No. 9, pp. 1686–1694 (2003)

[奥村 99] 奥村 学, 難波 英嗣：テキスト自動要約に関する研究動向, 自然言語処理, Vol. 6, No. 6 (1999)

[Skorochod'ko 72] Skorochod'ko, E.: Adaptive Method of Automatic Abstracting and Indexing, in *Proceedings of the IFIP Congress 71*, pp. 1179–1182 (1972)

[Uchida 00] Uchida, H., Zhu, M., and Senta, T. D.: UNL: A gift for a millennium, Technical report, The United Nations University (2000)

〔担当委員：西田 豊明〕

2004 年 10 月 13 日 受理

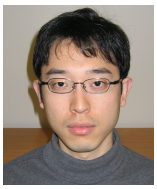
著 者 紹 介

綾 聡平

2002 年東京大学工学部電子工学科卒業．2004 年同大学院情報理工学系研究科電子情報学専攻修士課程修了．同年（株）東芝入社．現在，同社デジタルメディアネットワークス社にて，知的財産業務に従事．

松尾 豊（正会員）

1997 年東京大学工学部電子情報工学科卒業．2002 年同大学院博士課程修了．博士（工学）．同年より，産業技術総合研究所サイバーアシスト研究センター勤務． 2004 年 7 月より産業技術総合研究所情報技術研究部門．GBRC 社会ネットワーク研究所研究員．(株) ホットリンク技術アドバイザー．2002 年度人工知能学会論文賞受賞．最近は Web からの高次意味情報のマイニングに興味がある．受け手にとって価値の高い情報の提示を目指している．情報処理学会，AAAI の各会員．



岡崎 直観

2001 年東京大学工学部電子情報工学科卒業．2003 年同大学院情報理工学系研究科修士課程修了．現在同大学院博士課程在学中．2005 年より英国 NaCTeM のリサーチアシスタント．文書自動要約を中心にテキストマイニングの研究を行っている．電子情報通信学会，言語処理学会の学生員．

橋田 浩一（正会員）

1981 年東京大学理学部情報科学科卒業．1986 年同大学院理学系研究科博士課程修了．理学博士．1986 年電子技術総合研究所入所．1988 年から 1992 年まで (財) 新世代コンピュータ技術開発機構に出向．2001 年から産業技術総合研究所サイバーアシスト研究センター副研究センター長，ついで研究センター長．2004 年 7 月より産業技術総合研究所情報技術研究部門副部門長．専門は自然言語処理，人工知能，認知科学．現在の研究テーマはセマンティックコンピューティングおよびそれに基づく知の社会的共創など．

石塚 満（正会員）

1971 年東京大学工学部電子卒，1976 年同大学院博士修了．同年 NTT 入社，横須賀研究所勤務．1978 年東京大学生産技術研究所・助教授，(1980-81 年 Purdue 大学客員準教授)，1992 年東京大学工学部電子情報・教授，2001 年情報理工学系研究科電子情報学専攻，現在に至る．研究分野は人工知能，インターネット/WWW インテリジェンス，生命的エージェントによるマルチモーダルシステム．IEEE, AAAI, 情報処理学会，電子情報通信学会，映像情報メディア学会，画像電子学会，等の会員．