談話構造解析に基づくスライドの自動生成

柴田　知秀[†]　黒橋　禎夫[††]

本稿では，テキストから要約スライドを自動生成する手法を提案する．本稿で生成するスライドは，入力テキストから抽出したテキストの箇条書きからなる．それらに適切なインデントを与えるには，対比関係や詳細化関係などといった文または節間の関係を解析する必要がある．本手法では，まず，接続詞などの手がかり表現，語連鎖の検出，二文間の類似度の三つの観点を用いてテキストの談話構造を解析する．そして，テキストから主題部・非主題部を抽出・整形し，抽出したテキストのインデントを談話構造に基づいて決定することにより，スライドを生成する．実験を行なったところ，入力テキストよりもかなり見やすいスライドを自動生成できることが確認された．

キーワード:　談話構造解析，主題抽出，文簡約，自動プレゼンテーション

Automatic Slide Generation Based on Discourse Structure Analysis

Tomohide Shibata[†] and Sadao Kurohashi[††]

In this paper, we describe a method for automatically generating summary slides from a text. The slide consists of itemizations of extracted texts, and to determine their indentation, we need to analyze relations between sentences/clauses, such as contrast and elaboration. We first analyze the discourse structure of the text by considering three types of information: cue phrases, identification of word chain and similarity between two sentences. Then, we extract topic/non-topic parts from the text and generate the slide by placing the extracted texts, whose indentations are controlled according to the discourse structure. Our experiments demonstrate that generated slides are far easier to read in comparison with original texts.

KeyWords: *discourse structure analysis, topic extraction, sentence reduction, automatic presentation*

1　はじめに

スライドを用いたプレゼンテーションは，意見を人々に伝えるのに大変効果的であり，学会やビジネスといった様々な場面において利用されている．近年，PowerPoint や Keynote といったプレゼンテーションスライドの作成支援をするソフトが開発・整備されてきているが，一からスライドを作成することは依然として大変な作業である．

そこで，科学技術論文や新聞記事からプレゼンテーションスライドを自動 (または半自動) で生成する手法が研究されている．Utiyama らは，GDA タグで意味情報・文章構造がタグ付

† 東京大学大学院情報理工学系研究科, Graduate School of Information Science and Technology, University of Tokyo
†† 京都大学大学院情報学研究科, Graduate School of Informatics, Kyoto University

大阪と神戸を結ぶＪＲ神戸線，阪急電鉄神戸線，阪神電鉄本線の3線の不通により，一日45万人，ラッシュ時最大1時間12万人の足が奪われた．ＪＲ西日本東海道・福知山・山陽線，阪急宝塚・今津・伊丹線，神戸電鉄有馬線の不通区間については，震災直後から代替バスによる輸送が行われた．国道2号線が開通した1月23日から，同国道と山手幹線を使って，大阪～神戸間の代替バス輸送が実施された．1月28日からは，国道2号，43号線に代替バス優先レーンが設置され，効率的・円滑な運行が確保された．

図 1　入力テキストの例

鉄道の復旧

- 大阪と神戸を結ぶＪＲ神戸線，阪急電鉄神戸線，阪神電鉄本線の3線の不通
  - 一日45万人，ラッシュ時最大1時間12万人の足が奪われた
- ＪＲ西日本東海道・福知山・山陽線，阪急宝塚・今津・伊丹線，神戸電鉄有馬線の不通区間
  - 震災直後から
    - 代替バスによる輸送
  - 国道2号線が開通した1月23日から
    - 同国道と山手幹線を使って，大阪～神戸間の代替バス輸送が実施
  - 1月28日から
    - 国道2号，43号線に代替バス優先レーンが設置され，効率的・円滑な運行が確保

図 2　自動生成されたスライドの例

けされた新聞記事を入力としてプレゼンテーションスライドを自動生成している (Utiyama and Hasida 1999)．また，安村らは，科学技術論文のTEXソースを入力として，プレゼンテーション作成を支援するソフトウェアを開発している (安村禎明，武市雅司，新田克己 2003)．しかし，いずれの研究においても，入力テキストに文章構造がタグ付けされている必要があり，入力テキストを用意することにコストがかかってしまう．

本稿では，生テキストからスライドを自動生成する手法を提案する．入力テキストの例を図1，それから自動生成されたスライドの例を図2に示す．本稿で生成するスライドは，入力テキストから抽出したテキストの箇条書きから構成される．箇条書きを使うことによって，テキストの構造を視覚的に訴えることができる．例えば，インデントが同じ要素を並べることで並列/

対比関係を表わすことや，インデントを下げることによって詳細な内容を表わすことなどといったことが可能となる．従って，生成するスライドにおいて，箇条書きに適切なインデントを与えるには，入力テキストにおける，対比/並列関係や詳細化関係などといった文または節間の関係を解析する必要がある．本稿では，入力テキストの談話構造を解析し，入力テキストから抽出・整形されたテキストを箇条書きにし，そのインデントを入力テキストの談話構造に基づいて決定することによりスライドを生成する．生成されたスライドは入力テキストに比べて見やすいものにすることができる．特に，テキストに大きな並列や対比の構造があると，見やすいスライドを生成することができる．図2の例では，「震災直後から」，「国道2号線が開通した1月23日から」，「1月28日から」の対比の関係が解析され，それらが同じインデントで表示されることにより見やすいスライドが生成されている．また，図2の例の「震災直後から」と「代替バスによる輸送」のように，各文から主題を取り出し，主題部と非主題部を分けて出力することにより，スライドを見やすくしている．特に対比関係の場合，何が対比されているのかが明確になる．

本稿で提案するスライド生成の手法の概要を以下に示す．

(1)　入力文を JUMAN/KNP で形態素解析，構文解析，格解析する．

(2)　入力文を談話構造解析の基本単位である節に分割し，表層表現に基づいて談話構造解析を行なう．

(3)　入力文から主題部・非主題部を抽出し，不要部分の削除，文末の整形を行なう．

(4)　談話構造解析結果に基づき，抽出した主題部・非主題部を配置することによりスライドを生成する．

また，我々の手法は，プレゼンテーションスライドの作成支援を行なうだけでなく，自動プレゼンテーションを生成することができる．すなわち，テキストを入力とし，自動生成したスライドを提示しながら，テキストを音声合成で読み上げることにより，自動でプレゼンテーションを行なう．我々はこのシステムのことを，「text-to-presentation システム」と呼んでいる (図3)．難解な語や長い複合語は音声合成の入力に適しているとはいえないので，Kaji らの言い換え手法 (Kaji, Kawahara, Kurohashi, and Sato 2002; Kaji, Okamoto, and Kurohashi 2004) で書き言葉を話し言葉に自動変換してから音声合成に入力することにより，音声合成の不自然さを低減する．

本稿の構成は以下のようになっている．2章で談話構造解析について述べ，3章で入力テキストからスライドに表示するテキストを抽出する方法について述べ，4章でスライドの生成方法を述べる．そして，5章で実装した text-to-presentation システムと，自動スライド生成の実験の結果を報告する．6章で関連研究について述べ，7章でまとめとする．

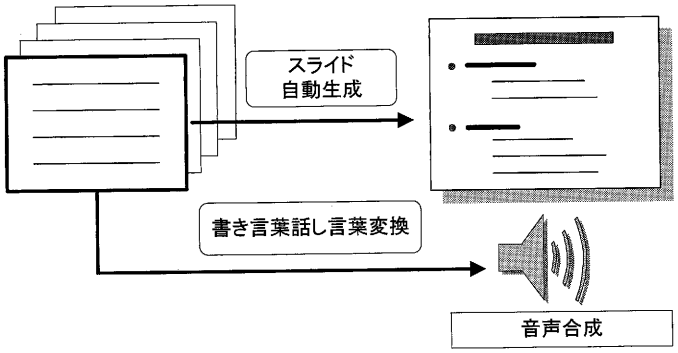

図 3　text-to-presentation システム

2　談話構造解析

この章では，テキストの談話構造を解析する手法を述べる．まず，談話構造のモデルを説明し，次に談話構造を解析する手順について説明する．

2.1　談話構造のモデル

談話構造を図 4 に示すようなものにモデル化した．図において，矢印，ラベルはそれぞれ，文 (S) または節 (C) の接続，結束関係を意味する．このモデルでは，初期状態として初期節点を設けており，文が初期節点に接続する時の結束関係を “初期化” とし，この文から新しい話題が始まることを意味する．

節と文を談話構造の基本単位とし，以下にあげる二種の結束関係を考える．どのような結束関係を考えるかは研究者によって異なるが，スライドを自動生成するためにはこれらで十分であると考えた．

(1)　一文内における節間の関係 (4 種類)
　　並列，対比，順接，逆接

(2)　二文間の関係 (11 種類)
　　並列，対比，理由，条件，主題連鎖，焦点主題連鎖，詳細化，理由，原因結果，例提示，質問応答

次節から，談話構造解析について述べる．解析の手順は (黒橋，長尾 1994) に基づいている．解析は入力文一文ずつ行ない，談話構造を逐次的に構築する．まず，入力文を節に分割し，節間の関係を解析する．次に，すでに入力した文の中で，入力文と最も関連する文とその間の結束関係を様々な手がかりをもとに決定する．図 4 の例は，1 文目から順番に談話構造を解析していき，4 文目までの構造が決定され，5 文目の解析を行なっている様子を示しており，様々な

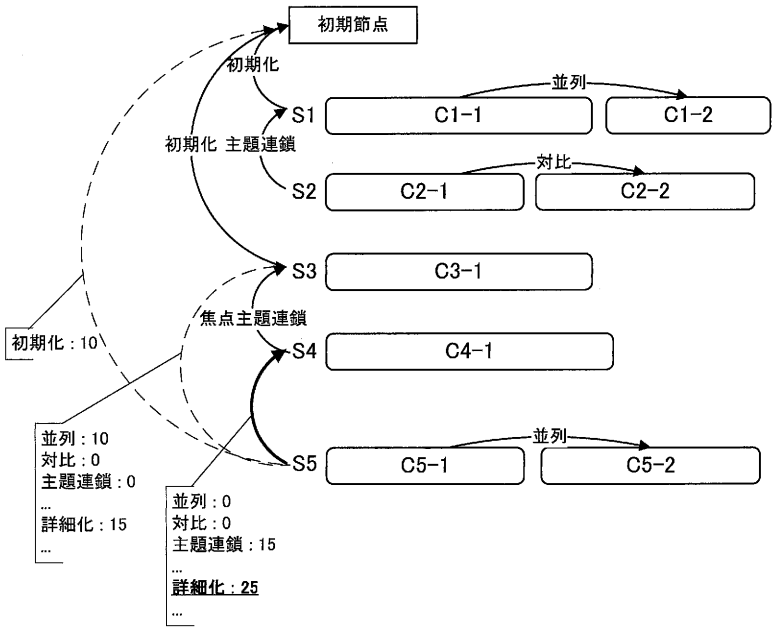

図 4　談話構造のモデル

文・結束関係の確信度を計算した結果，最も高い確信度を得た 4 文目と詳細化の関係で接続すると解析されている．

2.2　主題属性の付与

まず，入力文を JUMAN で形態素解析，KNP で構文・格解析を行なう．その後，2.4 節で述べる対比関係の抽出，2.5 節で述べる主題・焦点の抽出，3 章で述べる主題部・非主題部の抽出のために，あらかじめ，主題となりうる文節に主題属性を付与しておく．以下にあげるようなパターンを満たす文節に主題属性を付与する．パターンは形態素を単位に記述し，入力文の形態素解析結果と照合する．

- 延焼速度は，...
- インナーシティでは，...
- 出火原因の判明した火災において，...
- 3 線の不通により，...

また，以下のパターンの場合は～の部分に主題属性を付与する．

…{ する/した }{ の/とき } は～{ だ/になる }

以下の例では，「安否情報など」に主題属性を付与する．

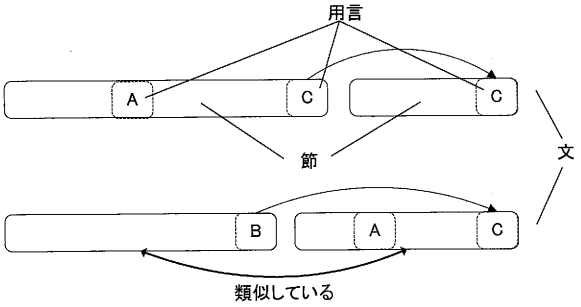

図 5　文の節への分割

(1)　震災直後に被災者が必要としたのは，家族や友人・知人の消息に関する安否情報などだった.

なお，ここで付与した主題属性を利用して，2.4 節や 2.5 節で対比・並列関係の解析を行なうが，関係を解析した結果，新たに主題属性が付与される場合がある.

2.3　入力文の節への分割

談話構造において，何をその基本単位とするかは研究者によって様々な定義がなされてきた. (Polanyi 1988; 黒橋，長尾 1994) では文，(Longacre 1983) では節，(Grosz and Sidner 1986; Marcu 1999b, 1999a) では独自に定義された単位 (clause-like unit) が談話構造の基本単位として採用されている．本研究では，スライドに配置する箇条書きが適切な長さとなるように，節に分割する基準を，南 (南不二男 1993) の従属節の分類に応じて以下のように設定した[1]．なお，節は談話構造解析の基本単位であると同時に，3 章で説明する主題部・非主題部の抽出の基本単位でもある.

レベル C　(例: ～が，～けれども) : 必ず分割する

レベル B　(例: ～て，～し) : 前後の文節列が類似している場合，または，節の文字数が閾値[2]以上の場合に分割する

レベル A　(例: ～ながら，～つつ) : 分割しない

レベル B における節の分割において利用している文節列の類似度は，KNP で並列構造を検出するために計算している任意の 2 つの文節列の類似度を用いる．任意の 2 つの文節列の類似度計算は，まず，あらゆる 2 文節の類似度を語の一致，品詞の一致，シソーラス (NTT コミュニケーション科学研究所 1997) による類似度などにより計算し，その上で，DP マッチングでそれらの文節間の類似度を組み合わせることにより行なわれる (Kurohashi and Nagao 1994).

1 本論文では，一つの述語からなるまとまりではなく，ここで定義したものを節と呼ぶことにする.
2 この閾値はスライドの横幅とフォントのサイズによって決定される.

(2) の例では,「達したが」のレベルがCなので節に分割し，(3) の例では「減り」のレベルがBで，前後の文節列が類似しているので節に分割する．(4) の例では，前後の文節列が類似していないので,「なく」で分割せず,「迫られて」もレベルがAなので分割を行なわない．

(2)　[神戸市には，他都市，業界等からの仮設トイレ支援が約3,000基に達したが，$_{\text{レベル } C}$] [受入れのための仮置き場の確保が大きな課題となった．]

(3)　[100人に1基行き渡った段階で設置についての苦情はかなり減り，$_{\text{レベル } B}$] [75人に1基達成できた段階では苦情が殆どなくなった．]

(4)　[震災時の環境保全については事前の具体的な対応策等がなく，$_{\text{レベル } B}$ 必要に迫られて $_{\text{レベル } A}$ 進めざるを得なかった．]

2.4　一文内の節間の関係の解析

入力文を節に分割した後，一文内の節間の結束関係を求める．まず，各節の親の節を構文解析結果に基づいて決定する．すなわち，各節の親を，節の最終文節の係り先の文節を含む節とする．次に，節間の結束関係を以下の基準で決定する．

- 二節が類似している場合
  - 並列
  - 対比
- 類似していない場合
  - 順接 (～て，(連用形))
  - 逆接 (～が，～けれども)

まず，二節が類似していない場合，結束関係を順接または逆接とし，順接であるか逆接であるかは節末の形態素列のパターンで認識する．

二節が類似している場合，結束関係を対比または並列とする．一般に並列の関係の場合，人または物がある二つの属性を持ち，対比の関係の場合，二つの異なる人または物が類似した属性を持つ．従って，二つの節において，主題属性が付与された二文節が，二節の類似度を計算する際のDPマッチングのパス上で対応関係にあり，かつ，それらの類似度が閾値以上である場合，結束関係を対比とし，そうでない場合を並列とする．

以下の例では，二節が類似しており，主題属性が付与された「当初は」と「3月末までは」が類似しているので，対比の関係とする．

(5)　[代替バス利用者は，当初は1日あたり3～5万人であったが，] [3月末までは1日約20万人が利用した．]

また，どちらか一方の節の文節に主題属性が付与されており，もう一方の節の文節には主題属性が付与されていない場合でも，類似度が高い場合は対比関係とする．以下の例では，二節が類似しており，主題属性が付与された，「(75 人に 1 基達成できた) 段階では」と主題属性のついていない「(100 人に 1 基行き渡った) 段階で」が類似しているので，結束関係を対比とし，「(100 人に 1 基行き渡った) 段階で」に主題属性を付与する．

(6)　[100 人に 1 基行き渡った段階で設置についての苦情はかなり減り，] [75 人に 1 基達成できた段階では 苦情が殆どなくなった．]

以下の例では，主題属性の付与された「パソコン通信ニフティサーブでは」と DP マッチングで対応付けられた「ボランティア情報」との間の類似度が閾値以下なので，結束関係を並列とする．

(7)　[パソコン通信ニフティサーブでは「地震情報コーナー」が開設され，] [ボランティア情報，安否情報，行政情報など各種の情報提供に用いられた．]

2.5　二文間の関係の検出

二文間の関係は，種々の表層的手がかりをもとに，各入力文に対して，関係をもつ以前の文 (接続文) とその間の結束関係を逐次的に求める．新しい話題が導入された後に古くなった話題に接続することはないという仮定をおき，入力文は談話構造の一番最後の子供の文にのみ接続可能と考える．図 4 では，文 5 は初期節点，文 3，文 4 に接続可能となり，文 1，文 2 との接続を許さない．そして，さまざまな接続可能文との間のさまざまな結束関係を考慮し，(1) 手がかり表現，(2) 語連鎖，(3) 二文の類似度の 3 つの観点から確信度を計算し，最終的に最も高い確信度を得た関係を採用する．以下，これらの 3 つの観点について順に詳しく述べる．

(1)　手がかり表現

種々の結束関係を示す，接続詞などの表層的な手がかり表現を認識し，その結束関係への確信度を得るために，表 1 に示すようなルールを用意した．表 1 において，接続可能文パターン，入力文パターンは，それぞれに対する表層表現，文間の結束関係 ([ ] で括られたもの) などのパターン，適用範囲とはどれだけ離れた文との関係まで考えるかである．適用範囲において，「1」は接続可能文と入力文が隣接している場合のみルールが適用されることを，「*」はルールの適用に制限がないことをそれぞれ意味する．ルールが一致した場合には，指定された結束関係欄の関係に対して，確信度欄の点数が与えられる．この確信度は経験的に決定した．

(2)　語連鎖の検出

一般に文は主題を示す部分 (主題部) とそれ以外の部分 (非主題部) に分けることがで

表 1　談話構造解析のルール

接続可能文パターン	入力文パターン	適用範囲	結束関係	確信度
～	さて～	*	初期化	10
～	そして～	1	並列	5
第一に～	第二に～	*	並列	30
[並列]	さらに～	1	並列	40
～	むしろ～	1	対比	30
～	すなわち～	1	詳細化	30
～	～からだ	1	理由	30

き，主題を 2.2 節で付与した主題属性のついている文節から名詞をとり出したもの，焦点を非主題部の名詞とする．そして，二文間で，主題と主題，焦点と主題に語の連鎖 (同一の語/句の出現) がある時は，それぞれ，主題連鎖，焦点主題連鎖の結束関係に確信度を与える．語連鎖は，完全一致と部分一致を考え，完全一致の場合は確信度 15 点を，部分一致の場合は確信度 10 点を与える．

(3)　二文間の類似度

二文が並列または対比の関係にある場合，それらはある種の類似性を持つと仮定することができる．二文間の類似度は，一文内の節の係り受け関係の所で述べた，任意の文節列間の類似度計算の方法で計算することができる．そして，一文内の対比/並列の関係の検出と同じように，二文における主題が類似している場合は対比の関係に，類似していない場合は並列の関係に確信度を与える．

以下の例では，二文が類似しており，主題属性が付与されていない文 (8-a) の「1 月 23 日から」と，主題属性が付与されている文 (8-b) の「1 月 28 日から」に高い類似が認められるので，対比の関係に確信度を与え，「1 月 23 日から」に主題属性を付与する．

(8)　a.　国道 2 号線が開通した 1 月 23 日から，大阪～神戸間の代替バス輸送が実施された．

b.　1 月 28 日からは，国道 2 号，43 号線に代替バス優先レーンが設置され，効率的・円滑な運行が確保された．

3　スライドに表示するテキストの抽出

この章では，入力テキストから，スライドに表示するテキストを抽出する手法を説明する．2.5 節で述べたように，文は主題部と非主題部から成る．入力テキストから文を抽出してそのままスライドに配置するのではなく，主題部と非主題部を分けてスライドに配置することにより，スライドを見やすいものとする．また，非主題部は一般に長いことが多いので，非主題部の簡

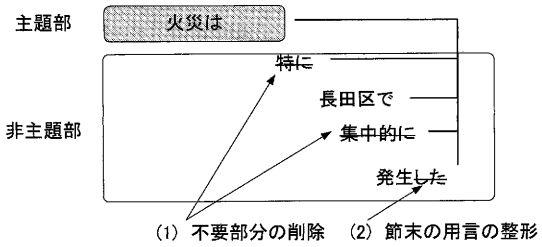

図 6　主題部・非主題部の抽出と非主題部の簡約

約を行なうことにより，スライドを見やすくする．主題部と非主題部の抽出は，2.3節で分割した節を基本単位として行なう．一連の解析の様子を図6に示す．

3.1　主題部・非主題部の抽出

2.2節で付与した主題属性をもとに，主題部の抽出を行なう．主題属性が付与された文節から構文木を子の方向にたどって句を抽出し，それを主題部とし，残りの部分を非主題部とする．以下の例では,「延焼速度」が主題部として抽出され,「おおむね 20〜40 m/h 程度で,」が非主題部となる．

(9)　延焼速度はおおむね 20〜40 m/h 程度で,

節に主題属性が付与された文節が複数存在する場合は，そのうち一番前にあるもののみを抽出する．以下の例では,「震災初日の被災地内では」と「視聴は」に主題属性が付与されているが,一番前の「震災初日の被災地内では」を主題部として抽出し，残りを非主題部とする．

(10)　震災初日の被災地内では停電などによりテレビの視聴はほとんどできず,

ただし，主題属性のついているもので，対比関係にあると解析されたものは必ず抽出する．以下の例では，前の節と後の節が対比の関係にあり,「神戸市では」に主題属性に主題属性が付与されており，主題属性の付与された「当初は」と「1月22日には」が対比の関係にあると解析されている．このような場合，前の節からは,「神戸市では」と「当初は」の両方を主題として抽出する．

(11)　[神戸市では，当初は 仮設トイレ 300 基程度で足りると考えていたが，] [1月 22 日には「仮設トイレ対策本部」を設置し対応することとなった．]

表 2　重要説明表現

格	用言
ガ格	重要だ，本質をつく，エッセンスだ，ポイント，望ましい，鍵だ，大切だ，有益だ，必要だ，指摘された
ヲ格	重視する，重要視する，明らかにする，明確にする，取り上げる
ニ格	着目する，重点を置く，注目する

3.2　非主題部の簡約

スライドを見やすいものとするためには，できるだけ入力テキストの情報を保持した上で，テキストを簡約する必要がある．本研究では，(1) 構文解析結果に基づく不要な語あるいは語句の削除，(2) 節末の用言の整形により，非主題部の簡約を行なう．

(1)　構文解析結果から不要部分の削除

構文解析結果から以下の不要な語句の削除を行なう．

- 接続詞
- 副詞
- レベル:A の節
- 副詞句

  例) バッテリー切れによる利用不能のほか，救援・復旧関係者による被災地外から大量持ち込みによる輻輳の発生で利用できなくなった．
- 同格：節末の用言の子の文節に「～など」があれば削除する．

  例) 農林水産省，国土庁など国の各機関

(2)　節末の用言の整形

次のようなルールにより節末の用言の整形を行なう．

- サ変名詞 + する/された → サ変名詞

  例) 実施された → 実施
- サ変名詞 + が行われた → サ変名詞

  例) 輸送が行われた → 輸送
- 名詞 + 判定詞 → 名詞

  例) 無被害であった → 無被害
- ナ形容詞/ナノ形容詞 → 活用語尾を削除

  例) 軽微であった → 軽微

ただし，節末の用言に否定表現を含む場合は，否定表現を削除してしまわないように，否定表現より後の部分の整形を行なう．

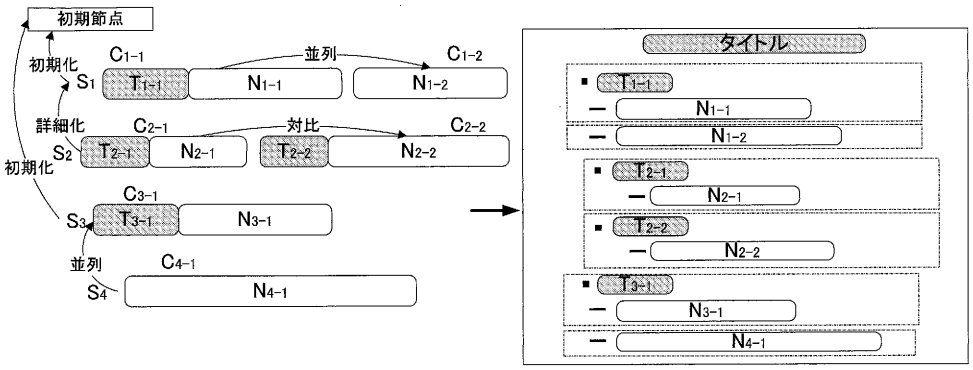

図 7　談話構造に基づくスライドの生成

3.3　強調表示

節末の用言が表 2 にあげるもので，かつ，指定された格を持つ場合，重要表現とみなし，次節で説明するスライドの出力の際に，この節の非主題部の強調表示を行なう．以下の例では，格解析の結果,「ことも」がガ格と解釈され,「指摘された」がガ格を持つことになるので，スライドの生成の際に強調表示を行なう．

(12)　大規模な供給施設が液状化地域に設置されていなかったことも指摘された．

4　スライドの生成

図 7 に示すように，2 章で解析した談話構造に基づいて，3 章で抽出した主題部，非主題部を次にあげるルールでスライドに配置する (図において，T，N はそれぞれ主題部，非主題部を表す)．

- メタデータなどで，入力テキストにタイトルがある場合，そのままスライドのタイトルとする．タイトルがない場合は，入力文の最初の主題部をスライドのタイトルとする．
- 文の先頭から順番に，各節において，主題部があれば出力し，インデントを一つ下げて次の行に，非主題部を出力する．主題部がなければ，非主題部だけを出力する．
  以下の例のように，節に主題が二つある場合は，一つ目の主題部を出力し，インデントを一つ下げて次の行に二つ目の主題部を出力し，さらにインデントを一つ下げて次の行に非主題部を出力する．

(13)　神戸市では，当初は仮設トイレ 300 基程度で足りると考えていたが，

⇓

- 神戸市
  - 当初
    - 仮設トイレ300基程度で足りると考えていた

また，3.3節の処理により，非主題部が重要表現とみなされている場合は，強調表示を行なう．

- 節のインデントのレベルを親との結束関係に応じて以下のように設定する．
  - **初期化:** インデントレベルを0にする．
  - **並列/対比:** 同じにする．
  - **主題連鎖:** 主題部が親と同じ場合は，インデントを下げずに非主題部だけを出力する．主題部が異なっている場合は，インデントを下げて，主題部と非主題部を出力する．
  - **その他:** 親に対してインデントを一つ下げて出力する．

出力される箇条書きの行数が閾値以上となる場合，各スライドの行数が閾値以下になるように分割し，複数のスライドを生成する．

また，多くの研究者が指摘しているように，一般に談話構造の根に近い方が文の重要度が高いと考えられるので，談話構造は要約生成のための手がかりとなりうる (Ono, Sumita, and Miike 1994; Marcu 1999b)．従って，根に近い文から抽出したテキストをスライドに出力し，談話構造木のある深さ以上の文から抽出したテキストはスライドに出力しないといった処理を行なうことにより，スライドに表示するテキストの量を制御することができる．しかし，自動生成したスライドを音声合成とともにユーザに提示する場合，音声に対応するテキストが全くないとユーザが違和感を感じてしまうので，上記の処理を行なわなかった．

5　実装と評価

5.1　text-to-presentation システムの実装

この節では，実装したtext-to-presentationシステムについて説明する．このシステムでは，ユーザは自然言語でクエリを入力すると，クエリに関するプレゼンテーションを閲覧することができる．本稿では，テキスト集合として，阪神淡路大震災教訓資料集[3]を用いた．この資料集はHTMLで書かれており，HTMLタグを手がかりとして400テキストに自動分割することによりテキスト集合とした．各テキストにはタイトルが付与されており，スライドのタイトルとして利用した．テキストの平均文数は3.7，一文あたりの平均文字数は50であった．

システムはまず，Kiyotaらの手法を用いて，ユーザからのクエリと最も類似したテキスト

3 http://www.hanshin-awaji.or.jp/kyoukun/

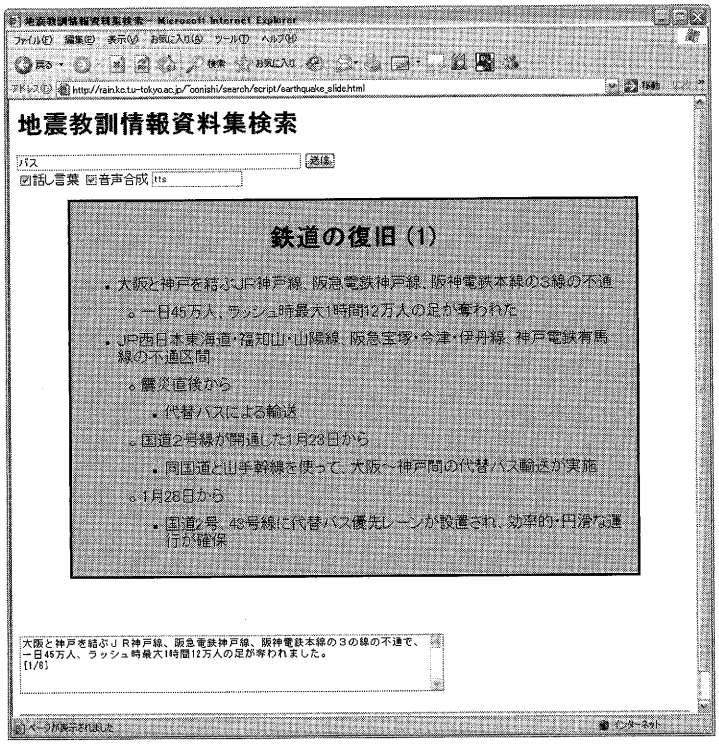

図 8　システムのスクリーンショット

を検索する (Kiyota, Kurohashi, and Kido 2002)．その後，書き言葉を話し言葉に変換して音声合成に入力し，同時に，本稿で述べた手法で生成したスライドをユーザに提示する．図 8 に示すように，本システムは Web ブラウザ上で動作する．テキストから複数のスライドが生成された場合は音声合成と同期してスライドの表示を切り換える．

5.2　評価と考察

「ボランティアの役割」「火災の原因にはどのようなものがありますか」などといったユーザからの 30 クエリから検索されたテキストからスライドを生成し，談話構造解析と生成されたスライドの評価を行なった．書き言葉からの話し言葉への変換とテキスト検索の評価に関しては，それぞれ (Kaji et al. 2002, 2004)，(Kiyota et al. 2002) に譲る．入力テキストと自動生成されたスライドのサイズを比較した平均圧縮率は 0.797 であった．

表 3　談話構造解析の精度

	精度
節間の関係	30 / 39 (76.9%)
文間の関係	60 / 89 (67.4%)

5.2.1　談話構造解析の評価

談話構造解析の精度を表3に示す．評価は，節または文の接続先と結束関係が正しいかで行なった．談話構造解析の主な誤り原因を以下に示す．

語連鎖の検出もれ　以下の例では1文目の「震災」と2文目の「地震」の関係が捉えられず，2文目で初期化されてしまっている．

(14)　a.　震災直後には，神戸市によって神戸市外語大のホームページに被害写真が掲載され，海外に被害の大きさを知らせた．
　　b.　パソコン通信ニフティサーブでは「地震情報コーナー」が開設され，ボランティア情報，安否情報，行政情報など各種の情報提供に用いられた．

この問題には，国語辞典やシソーラスを用いて表現のずれを認識することにより対処することができると考えられる．

また，本稿で扱ったテキストは書き言葉のため，語が省略されているために語連鎖が捉えられない例はそれほどなかったが，以下のような例では，「延焼速度」には「火災の」が省略されており，この関係が捉えられていないため，二文間の関係を主題連鎖と解析することができず，初期化されてしまった．

(15)　a.　このうち焼損面積 10,000 平方メートル以上の火災は，特に神戸市長田区などで集中的に発生した．
　　b.　延焼速度はおおむね 20～40 m/h 程度で，過去の都市大火事例等と比較して極めて遅かった．

この問題には省略・照応解析を行なうことで対処する予定である．

対比関係の検出もれ　名詞と句/節などが対比の関係にある時に，対比関係を検出できないことがあった．以下の例では，「震災直後に」と「震災から1週間程度を経ると」の対比関係を検出できなかった．

(16)　a.　震災直後に被災者が必要としたのは，地震の規模や発生場所，被害状況などの被害情報，家族や友人・知人の消息に関する安否情報などだった．

b. 震災から1週間程度を経ると，長期的な生活に関わる情報として，住宅やり災証明を始めとする各種申請などの情報も求められた．

また，以下の例では，「当初の」，「時間とともに」の対比関係が検出できなかった．

(17) ボランティアの当初の役割は，医療，食糧・物資配給，高齢者等の安否確認，避難所運営等だったが，時間とともに，物資配分，引っ越し・修理，高齢者・障害者のケアなどへと変化していった．

この問題には，「時間とともに」や「震災から1週間程度を経ると」が時間に関する表現であることを認識するためのルールを用意した上で，節/文の類似度を計算することで対処することができると考えられる．

また，本稿で対象とした地震ドメインのテキストにはシソーラスにない語が含まれており，語の類似度を正しく計算できないことがあった．それが原因となり，文/節の類似度を正しく計算できないことがあった．この問題には地震ドメインにおけるシソーラスを人手で用意するか，コーパスから自動構築することにより対処できると考えられる．

5.2.2 自動スライドの出力例と評価

次に，自動生成したスライドの評価を行なった．評価の基準は，生成されたスライドがユーザの理解を妨げるものとなっていないかどうかとし，スライドのインデント，主題の抽出，文簡約などが適切であるかどうかをもとに評価を行なった．生成した30スライドについて筆者らが評価したところ，15枚については自然であり，12枚は少し不自然なところが含まれており，3つは全体的に不自然であるという結果であった．出力例を図9と図10に示す．図9の例では，「断水」，「停電」，「都市ガスの供給停止」の対比関係が正しく解析され，また，「明かりに不自由しながらの診察・治療が行われ」と「手動の人工呼吸器を押し続ける姿も見られた」の並列関係で項目を分割することにより，見やすいスライドとなっている．また，2文目の「医療用水のほか」の簡約や，「(治療) が行なわれ」の整形などが行なわれている．

また，図10の例では，1文目の主題「代替バス利用者」が抽出され，「当初」，「バスレーン設置後」，「3月末」の対比関係が正しく解析されている．しかし，2文目では，「代替バス」と「バスレーンの設置後」が対比していると間違って解析されており，正しくは，「当初」と「バスレーンの設置後」が対比関係にある．この対比関係を正しく解析できるようにした上でさらにこのスライドをよくするには，2文目からは「利用時間」という主題を取り出し，1文目の「代替バス利用者」と対比させるのが望ましいが，これはかなり難しい処理であるといえる．

断水により，水の調達に苦慮した医療機関が多かった．断水の影響には，医療用水のほか，ボイラー用水や，コンプレッサー・自家用発電機等の冷却水が得られないという面もあった．停電により，明かりに不自由しながらの診察・治療が行われ，手動の人工呼吸器を押し続ける姿も見られた．都市ガスの供給停止により，入院患者の食事提供に影響があった病院もある．

⇓

被災地医療機関

- 断水
  - 水の調達に苦慮した医療機関が多かった
  - 断水の影響
    - ボイラー用水や，コンプレッサー・自家用発電機等の冷却水が得られないという面もあった
- 停電
  - 明かりに不自由しながらの診察・治療
  - 手動の人工呼吸器を押し続ける姿も見られた
- 都市ガスの供給停止
  - 入院患者の食事提供に影響があった病院もある

図 9　出力例 1

自動生成されたスライドで不自然な部分のほとんどは，談話構造解析誤りによるインデントのずれによるものであった．4章で説明したスライド生成のヒューリスティックルールによる大きな誤りはなく，また，主題部・非主題部の抽出や非主題部の簡約にも大きな誤りは見受けられなかった．現在の文簡約は構文解析結果を基に行なう比較的シンプルなモデルであるが，十分機能しているといえる．しかし，以下のように固有表現にかかる連体節などは削除した方がよりよいスライドになると考えられるので，今後は固有表現抽出を行ない，このような簡約を行なう予定である．

(18)　大阪と神戸を結ぶ ＪＲ神戸線，阪急電鉄神戸線，阪神電鉄本線 の 3 線の不通により，一日 45 万人，ラッシュ時最大 1 時間 12 万人の足が奪われた．

(19)　兵庫県下随一の 3 次救急医療機関である 神戸市立中央市民病院 は，市街地と島を結ぶ神戸大橋の不通により震災直後の救急患者の受け入れがあまりできなかった．

たとえ自動スライドにインデントのずれや抽出したテキストが不自然であるといった誤りが少

代替バス利用者は，当初は1日あたり3〜5万人であったが，バスレーン設置後は上昇し，3月末までは1日約20万人が利用した．当初，代替バスは交通渋滞に巻き込まれ，通行に多くの時間を要したが，バスレーンの設置後は約半分の所要時間に短縮されるなど，徐々に時間は短縮された．

⇓

鉄道の復旧

- 代替バス利用者
  - 当初
    - ＊ 1日あたり3〜5万人
  - バスレーン設置後
    - ＊ 上昇
  - 3月末
    - ＊ 1日約20万人が利用
- 代替バス
  - 交通渋滞に巻き込まれ，通行に多くの時間を要した
- バスレーンの設置後
  - 約半分の所要時間に短縮されるなど，時間は短縮

図 10　出力例 2

しあったとしても，入力テキストを音声合成と自動スライドのマルチモーダルに変換することは，ユーザに入力テキストをそのまま提示するよりもはるかによいことが実験により示された．特に，テキストに大きな並列や対比関係がある場合は，入力テキストよりも見やすいスライドを生成できることが確認された．

6　関連研究

Utiyama らは，GDA で意味情報・文章構造がタグ付けされた文書からスライドショーを生成する手法を提案している (Utiyama and Hasida 1999)．GDA タグとは，文書に意味論的構造や語用論的構造を与えるもので，人手で付与される．まず，共参照を示すタグから文章構造をボトムアップに決定する．そして，重要なトピックを抽出し，各トピックに対して関連する文を集め，それらを箇条書きにして一枚のスライドを生成する．GDA タグを用いることにより，ある程度長い文章についても文章構造を解析し，スライドを生成することができるが，GDA タ

グを付与するコストは大きなものとなる．

安村らは，論文からプレゼンテーション資料の作成支援を行なっている (安村禎明他 2003)．まず，論文中の各セクションに対して，使用するスライドの枚数を割り当て，そして，個々のスライドに対してレイアウトを決定し，論文中から抽出した文や図表といったオブジェクトを配置している．しかし，この研究では TF*IDF 法で重要文を抽出しており，文章構造の解析や文簡約は行なわれていない．また，入力は TEX 形式の論文に限られており，本研究のように生テキストからスライドを生成することができない．

次に，個別の処理に関連する研究をあげる．まず，談話構造解析の分野でよく知られているものとして，Marcu らの研究がある (Marcu 1999a, 2000; Carlson, Marcu, and Okurowski 2001)．彼らは談話構造タグ付きのコーパスを作成し，機械学習の手法を用いることにより談話構造解析を行なっている．彼らの手法には精度の向上が見られるが，談話構造タグ付きコーパスを作成するにはコストがかかってしまう．これに対して，我々の談話構造解析は一般的なヒューリスティックルールに基づいている．我々のシステムの確信度などはもともと比較的少数の科学技術文章を対象に経験的に定めたものであるが，そのままの設定で地震ドメインのテキストに対しても，スライドを生成するのに十分な精度を達成しているといえる．従って，地震ドメインで談話構造タグ付きコーパスを作成し，機械学習を行なう必要はないと考えている．

また，文末表現の整形で関連するものとして，(山本和英，池田諭史，大橋一輝 2005) の研究がある．この研究では，体言止めや助詞止めといった文末表現に着目し，新聞記事の表現を，新幹線車内や街頭での電光掲示板で流れるニュースで使われる表現に変換する手法を提案している．手法は我々と同じルールベースで，本研究で扱っているものよりも多くのパターンを利用しているが，誤り例も報告されており，我々の扱ったパターンでも十分であると考えている．

Jing らは，自動要約の質を向上させるために，新聞記事とそこから作られた人間による要約のペアから文簡約の手法を学習している (Jing 2000)．本研究においても，論文とプレゼンテーションスライドのペアから文の対応関係をとる研究 (Hayama, Nanba, and Kunifuji 2005) を利用して，このアイデアを適用し得ると考えられる．

7　おわりに

本稿では，テキストからスライドを自動生成する手法について述べた．スライド生成は，談話構造解析，主題部と非主題部の抽出と簡約によるスライドに出力するテキストの抽出，抽出したテキストの適切配置からなる．地震教訓集を入力テキストとして実験を行なったところ，生成されたスライドは入力テキストよりもかなり見やすいものであることが確認された．また，テキストを入力として自動プレゼンテーションを行なう，text-to-presentation システムの実装を行なった．

今後は，談話構造解析，主題の抽出，文簡約などの精度を高めるとともに，実装した text-

to-presentation システムに会話エージェントを統合しシステムの質を向上させる予定である．また，システム全体が自然なプレゼンテーションであるかや，ユーザの理解の向上に貢献するかについては今後，評価実験を行なう予定である．

参考文献

Carlson, L., Marcu, D., and Okurowski, M. E. (2001). "Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory." In *Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue.*

Grosz, B. J. and Sidner, C. L. (1986). "Attention, intentions, and the structure of discourse." *Computational Linguistic*, **12**, pp. 175–204.

Hayama, T., Nanba, H., and Kunifuji, S. (2005). "Alignment between a Technical Paper and Presentation Sheets Using Hidden Markov Model." In *Proceedings of the 2005 Internatinal Conference on Active Media Technology.*

Jing, H. (2000). "Sentence Reduction for Automatic Text Summarization." In *Proceedings of the sixth conference on Applied natural language processing*, pp. 310–315.

Kaji, N., Kawahara, D., Kurohashi, S., and Sato, S. (2002). "Verb Paraphrase based on Case Frame Alignment." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 215–222.

Kaji, N., Okamoto, M., and Kurohashi, S. (2004). "Paraphrasing Predicates from Written Language to Spoken Language using the Web." In *Proceedings of the Human Language Technology Conference*, pp. 241–248.

Kiyota, Y., Kurohashi, S., and Kido, F. (2002). "Dialog Navigator: A Question Answering System based on Large Text Knowledge Base." In *Proceedings of 19th COLING*, pp. 460–466.

Kurohashi, S. and Nagao, M. (1994). "A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures." *Computational Linguistics*, **20** (4).

Longacre, R. (1983). *The Grammar of Discourse.* New York: Plenum Press.

Marcu, D. (1999a). "A decision-based approach to rhetorical parsing." In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 365–372.

Marcu, D. (1999b). "Discourse trees are good indicators of importance in text." In I.Mani and M.Maybury (Eds.), *Advances in Automatic Text Summarization*, pp. 123–136. The MIT Press.

Marcu, D. (2000). "The Rhetorical Parsing of Unrestricted Texts: A Surface-Based Approach." *Computational Linguistics*, **26** (3), pp. 395–448.

NTT コミュニケーション科学研究所 (1997). “日本語語彙大系.” 岩波書店.

Ono, K., Sumita, K., and Miike, S. (1994). “Abstract generation based on rhetorical structure extraction.” In *Proceedings of the 15th COLING*, pp. 344–348.

Polanyi, L. (1988). “A formal model of the structure of discourse.” *Jounnal of Pragmatics*, **12**, pp. 601–638.

Utiyama, M. and Hasida, K. (1999). “Automatic Slide Presentation from Semantically Annotated Documents.” In *1999 ACL Workshop on Coreference and Its Applications.*

安村禎明，武市雅司，新田克己 (2003). “論文からのプレゼンテーション資料の作成支援.” 人工知能学会論文誌, **18** (4), pp. 212–220.

山本和英，池田諭史，大橋一輝 (2005). “「新幹線要約」のための文末の整形.” 自然言語処理, **12** (6), pp. 85–111.

黒橋禎夫，長尾眞 (1994). “表層表現中の情報に基づく文章構造の自動抽出.” 自然言語処理, **1** (1), pp. 3–20.

南不二男 (1993). 現代日本語文法の輪郭. 大修館書店.

略歴

柴田 知秀: 2002 年東京大学工学部電子情報工学科卒業．2004 年東京大学大学院情報理工学系研究科修士課程修了．現在，東京大学大学院情報理工学系研究科博士課程在学中．自然言語処理の研究に従事．

黒橋 禎夫: 1994 年京都大学大学院工学研究科電気工学第二専攻博士課程修了．博士（工学）．2006 年 4 月より京都大学大学院情報学研究科教授．自然言語処理，知識情報処理の研究に従事．

(2005 年 11 月 4 日 受付)
(2006 年 1 月 25 日 再受付)
(2006 年 2 月 26 日 採録)