# テキストの概要把握支援のための話題構造抽出

竹下 敦[†] 井上孝史[††] 田中一男[††]

ユーザがテキストの概要の直感的な把握を容易に行えるような環境を実現するために，話題構造を自動的に抽出する方法を提案する．我々の方法は，話題構造を読み手にうまく伝えるために，書き手が意識的あるいは無意識に用いる言語現象を規則化することにより，話題構造を抽出する．言語現象としては，タイトルや章立て，段落のような文書の論理構造，「まず」や「次に」などの手掛かり句，疑問表現，段落の長さ，「に関して」や「が」のような話題マーカ，連体修飾関係，固有名詞などの品詞情報を用いた．これらは非常に汎用的であるので，我々の方法は新聞記事や技術文書はいうまでもなく，電子メイルなどの多種多様なテキストにも適用することができる．現実のテキストに対して人間と我々のシステムが抽出した話題構造を比較する評価実験を行った結果，構造を考慮した話題スコープの評価で再現率が59.4%，適合率が59.1%であった．これは，概要把握という目的のために人間が利用できる精度である．さらに，現実のテキストを扱う際に避けられない未知語の問題に対しても検討を行い，1つのテキストに含まれる未知語の数が多くない場合は，我々の方法は精度を落とさないという見通しも得た．

# Text Topic Structure Extraction for a Skimming Interface

ATSUSHI TAKESHITA,[†] TAKAFUMI INOUE[††] and KAZUO TANAKA[††]

This paper proposes a new method for automatic topic structure extraction. The topic structures allow users to skim texts, and thus select necessary texts efficiently. Our method incorporates the linguistic phenomena that writers use consciously or unconsciously for communicating topic structures to readers: document logical structures such as a title, chapters and paragraphs, cue phrases like "mazu (first)" and "tsugi ni (next)", interrogative expressions, topic markers including "ni kanshite (with regard to)" and "wa (as for)", embedded sentences, and parts of speech. These very general phenomena make our method applicable to a variety of texts including newspaper articles, technical documents and e-mail messages. An experimental evaluation was performed by comparing human and system topic structures. Recall and precision ratios for topic structures were 59.4% and 59.1% respectively, which means that the extracted structures are acceptable for human users for the purpose of text skimming. Furthermore, the influence of the inevitable problem of unknown words was investigated. The investigation shows that the problem does not decrease the ratios of recall or precision if a text includes one or two unknown words.

## 1. はじめに

インターネットにおけるWorld Wide Webなどの情報提供環境や電子メイル，電子ニュースの浸透，オフィスにおける文書共有型グループウエア環境の普及など，個人がテキスト情報を発信する文化が定着してきた．しかしながら，現状では一方的な情報発信が行われているだけであり，情報の受け手がそれらを取捨選択し，活用するための環境は不十分である．

全文検索やフィルタリングなどの技術は，ユーザが与えた検索条件に合致する情報を選んでくれるが，検索精度には限界があるので，最終的にはユーザが検索結果のテキストを1つ1つ吟味して，取捨選択する必要がある．また，電子メイルのように情報を受動的に受け取る場合や，検索条件が明確には決まってない情報散策の場合は，全文検索やフィルタリングはそぐわないので，人手による情報選択が主たる手段となる．

人手による情報選択を支援するための技術として，要約の自動作成の研究，開発が行われてきたが，実用的なシステムはできていない．最近では，科学技術論文など対象を限定し，それらに特徴的に現れる言語表現などを用いて要約を自動作成する実用的技術も提案されている[1)]が，多種多様なテキストが発信されてい

† NTT北海道法人営業本部
NTT Hokkaido Business Communications Headquarters
†† NTTヒューマンインタフェース研究所
NTT Human Interface Laboratories

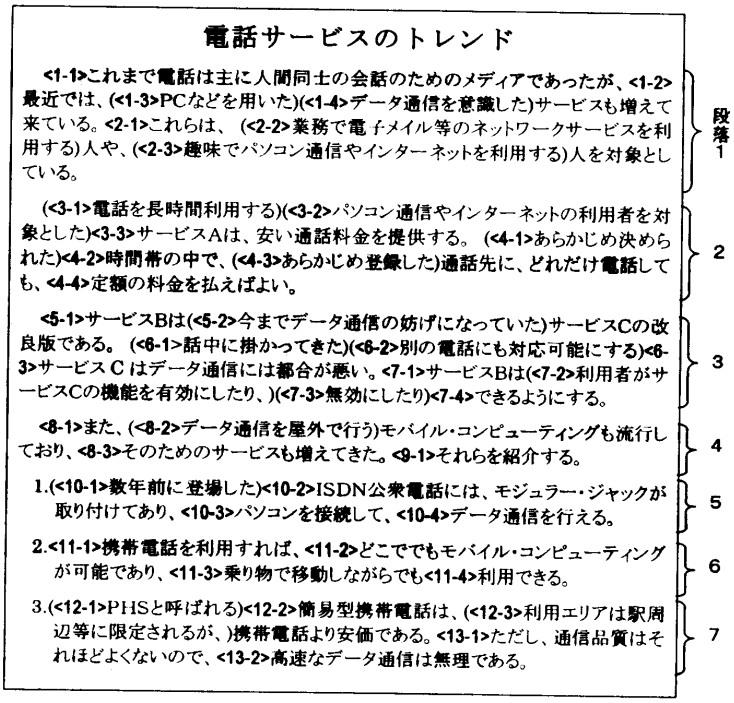


図1 テキスト例
Fig. 1 A text example.

る現在の状況では，より広範囲なのテキストに対して適用できる技術が望まれる．

本論文では，人手によるテキスト情報選択支援のための，実用的で適用範囲の広い話題構造抽出方法を提案する．テキストから自動抽出された話題構造によって，ユーザは概要を直感的に把握するだけでなく，熟読の必要性や緊急性を判断できるようになる．

まず，我々が提案する話題構造抽出方法の考え方について述べ，次に具体的なアルゴリズムを説明し，最後に評価実験の結果について検討する．

## 2. 話題構造抽出の考え方

### 2.1 話題構造の定義

我々の目的は，様々な内容のテキストの概要をユーザが把握することを支援する現実的な方法を確立することである．したがって抽出する構造は，読み手に対してその内容を直感的に伝達できものが望ましい．F. Danesによって提案された主題進行パターン[2]はテキストの内容を反映しているという点では適しているが，明示的に述べられていない上位概念を主題として抽出するという深い理解も行う必要があるので，現実的な手法を構築することが困難である．

我々は人間にテキストを与えて，「同じことが書いてあるブロックと，テキスト中の名詞句で，その『同じこと』を表すものを求めよ」という課題を与えたとき，回答として得られる構造を「話題構造」と呼び，これを抽出対象とした．話題構造は，この定義の『同じこと』に相当し，何に関するものかを示す「話題」と，そのブロックがどの文からどの文まで継続するかという「話題スコープ」によって表現できる．また，話題スコープの包含関係によって生じる話題の入れ子関係は「話題レベル」として扱い，一番外側の話題を1，入れ子の内側ほど1ずつ増加させるものとする．

図1のテキストに対する話題構造を図2に示す．このテキストでは，述語を1つだけ含む単位である単文ごとに，〈1-1〉のような番号が付与されているが，前の数字はテキスト全体における文番号を，後は各文における単文番号を示す．また，単文の範囲を示すために，他の単文中に埋め込まれた単文を“(”と“)”で囲んだ．図2において，話題「データ通信」のスコープは文〈1〉の文頭から〈7〉の文末までであり，話題レベルは1である．話題「サービスA」は〈3〉から〈4〉までのスコープを持ち，レベルは2である．

### 2.2 話題構造抽出の着眼点

我々は適用範囲の広い話題構造方法を目指すため，話題導入の際の言語現象に着目した手法を提案する．書き手は話題構造を読み手にうまく伝達するために意識的あるいは無意識に「次に」のような言語的手掛かりを用いるはずである．我々は，話題構造抽出のために，これらの言語現象を規則化した．

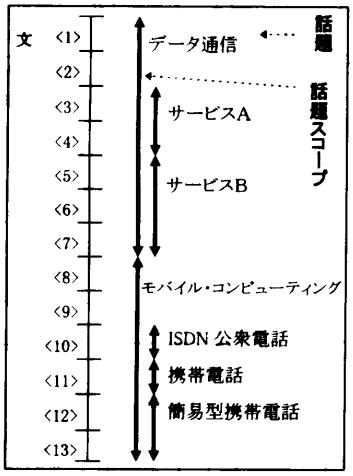


図2 人間による話題構造抽出の例
Fig. 2 An example of manual topic structures.

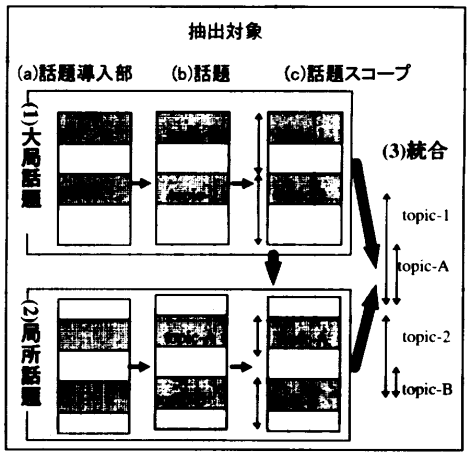


図3 話題構造抽出の流れ
Fig. 3 Outline of topic structure extraction.

話題構造抽出の流れを図3に示す.「次に」のような明示的手掛かりを用いて，大局的な話題構造が伝達されるが，これを「大局話題」と呼ぶことにする．手掛かり句などが多用されると，対応を取るのが困難になるので，明示的手掛かりなしに，より小さな話題が展開されるが，これを「局所話題」と呼ぶことにする．たとえば，段落という単位は話題構造には対応しないが，局所話題の手掛かりとして利用できる．抽出処理としては，まず大局話題を抽出し，次に局所話題を抽出し，最後に両者を統合する．最初の2つの各抽出処理では，話題導入部の抽出による処理対象文の絞り込み，話題の抽出，話題スコープの決定を順に行う．

我々の方法を従来の研究と比較する．最も関連深いものは，F. Danes による主題進行パターンを認識するための U. Harn らの方法である[2]が，対象領域の知識を用いる知識工学的アプローチであるので，適用範囲が非常に限定されており，我々の目的には合わない．山本らは手掛かり句を用いて，段落のない文章を自動的に段落分けする方法を提案した[3]．手掛かり句という点では，我々の大局話題の処理と類似するが，次の2点で異なる．1つ目は，我々の方法は意味的な切れ目だけでなく，話題とその入れ子構造まで抽出する点であり，2つ目は我々の方法は段落を抽出対象とはせず，むしろ手掛かりとして用いるという点である．

### 2.3 着目する言語現象

まず，話題構造の手掛かりとして着目した言語現象をリストアップする．

(1) 表題，章立て，箇条書き，段落などの文書論理構造：書き手が明示的に与えた構造であり，最大の手掛かりである．段落は話題構造と対応はしていないが，関係はある．たとえば，長い段落では子話題が提示されやすい．

(2) 「次に」などの手掛かり句：話題の導入だけでなく，話題の入れ子関係も示す．

(3) 「尋ねる」のような疑問表現：子話題を提示しやすい．

(4) 「について」や「が」などの話題マーカ：マーカによって話題の提示しやすさに差がある．

(5) 連体修飾関係：連体修飾されている語句は，話題として選ばれやすい．

(6) 固有名詞：指示対象が決まっており意味が明確であるので，話題として選ばれやすい．

次に，上記の言語現象を生じさせる要因をあげる．

- **要因1—記録性：** 書き言葉のテキストでは文字が消えない．
- **要因2—インタラクションのなさ：** 相手に割り込まれることなく，テキストを書くことができる．電子メイルでもこれは成り立つ．
- **要因3—話題を伝えたいという意志：** 情報伝達を行う言語データすべてに成り立つ．

最後に，言語現象と要因の対応付けを行う．(1)の論理構造は要因1によって可能となったが，テキストを分かりやすく表現するために用いられる．(5)の連体修飾関係は，要因1と要因2によって多用できるようになった．すなわち，読み手が読む速度を自分で調整でき，かつインタラクティブに説明する必要がないので，複雑な連体修飾関係を含む文が許される．(3)の疑問表現は要因2にそぐわないためにテキスト中で目立つ．対話で質問によって話題導入が行われることがある[4]が，それと類似している．また，要因3は(1)〜(6)のすべての言語現象の根本にある．

この対応付けにより，我々の方法が適するテキスト

表1 手掛かり句
Table 1 Cue phrases.

| | |
|---|---|
| 入り子開始型 | まず第一に，最初の，第1に |
| 話題転換型 | ところで，それから，さて，さっそく，このあと，そして，次に，次は，続く，第2に，……，第9に，第二に，……，第九に，こうしたことから，それに対し，それによると，むしろ，このほか，これまでに，そのうえで，その際 |
| 入り子終了型 | 最後，終わりに |

表2 明示マーカ
Table 2 Explicit topic markers.

| |
|---|
| に関して，について，として，とは，というのは，といえば，というと，という，とゆう，といったら，とくると，ときたら，となると，となれば，になると，となっては，に至ると，に至っては，かというと，かといえば，としては，としてみれば，としてみては，としても，の場合，にしても，にしたって，にしろ，にしては，といっても，といえども，におかれましては，ということ，なのです，の事で，のことで，によると，は |

の種類が明確になる．たとえば，新聞記事，論文，電子メイルなどには適する．記録性の少ない電光掲示板のテキストは適さない一例であり，これに対処するためには，要因1に起因する(1)論理構造と(5)連体修飾関係の処理規則を改良すればよい．

## 3. 話題構造抽出アルゴリズム

### 3.1 大局話題の導入部の抽出

第1に，導入部の候補をタイプごとに抽出する．1つ目はリード・タイプである．テキストの先頭にリード，すなわち章に含まれない文章がある場合はその第1文を，章立てがない場合はテキストの第1文を候補とする．2つ目は章立てタイプである．各章の第1文を候補とする．ただし，章タイトルAの直後に下位の章タイトルBが連続する場合は，Aを候補とする．

3つ目は手がかり句タイプである．「まず」「第1に」など，新しい子話題を導入する入れ子開始型，「次に」「また」など，前の子話題と並立した子話題を導入する話題転換型，「最後に」など，最後の子話題を提示する入れ子終了型の手掛かり句をあらかじめ辞書に登録しておき，テキストの最後以外の単文に手掛かり句が含まれていれば，その文を候補とする．現在，表1に示す手掛かり句が登録されている．

4つ目と5つ目は箇条書き全体タイプと箇条書き項目タイプであるが，それぞれ全体，項目タイプと略記することもある．人間による話題構造抽出実験でのインタビューによると，箇条書きにともなう話題が抽出されるかどうかは，各項目の長さに影響される．したがって，最後以外の各項目に，単文がいくつ含まれているかを調べ，もし，あらかじめ与えた値 *item-size-1* 以上であれば，箇条書きの直前の1文を全体タイプの候補とし，別の値 *item-size-2* 以上であれば，各項目の第1文を項目タイプの候補とする．

図1のテキストでは，リード・タイプの候補として文〈1〉が，手掛かり句タイプとして文〈8〉が抽出される．*item-size-1* と *item-size-2* の各値を4と3と仮定すると，全体タイプの候補として文〈9〉が，項目タイプとして文〈10〉,〈11〉,〈12〉が抽出される．なお，両パラメータの値とその決定方法は4.2節で説明する．

第2に，各話題導入部候補に含まれる各単文の中で，最も強調されている名詞句を顕著名詞句として抽出する．強調の度合いはセンタリングと同様に，名詞句を提示するマーカの優先順位によって決定し[5]，代名詞のように単独では意味を表さない語以外の名詞句で，最高順位のマーカで示されたものを選ぶ．「について」「は」のように話題を提示するための語句を明示マーカ，格助詞の「が」「を」のように主語や目的語などの文法機能を示す語句を非明示マーカと呼ぶと，優先順位は以下の順になる．現在，システムに登録されている明示マーカは表2に示すとおりであり，非明示マーカとしては13語が登録されている．

1.「は」以外の明示マーカ＋読点「、」
2.「は」以外の明示マーカ
3.「は」「は、」
4. 非明示マーカ

ただし，例外として，章タイトル自身が話題導入部候補となっており，これらの話題マーカがそこに含まれない場合は，章タイトル自身を顕著名詞句とする．

図1のテキストでは，単文〈1-1〉からは「電話」が明示マーカ「は」によって，〈1-2〉からは「サービス」が非明示マーカ「も」によって，〈1-3〉と〈1-4〉からは「PCなど」と「データ通信」が非明示マーカ「を」によって顕著名詞句として抽出される．〈9-1〉では非明示マーカ「を」によって提示される名詞句が，代名詞「それら」であるので，顕著名詞句は抽出されない．他の単文についても同様の処理が行われる．

第3に，顕著名詞句を含まない導入部候補を拡張する．全体タイプは前方に，それ以外のタイプは後方に，顕著名詞句が見つかるまで拡張する．ただし，顕著名詞句が見つかる前に，隣の候補に到達した場合は，以下の規則に従う．すなわち，両候補とも章立てタイプ

か項目タイプのいずれかである場合は，拡張を中止する．その候補からは後の処理でも話題は抽出されないので，「空話題」として扱う．それ以外の場合は以下の優先順位で，低いタイプの候補を棄却する．

章立て，箇条書き項目 > リード > 手掛かり句 > 箇条書き全体

また，最初から候補が重なっている場合も，この優先順位が低い方を棄却する．

図1のテキストでは，文〈9〉の候補は顕著名詞句を含まず，かつ全体タイプであるので，前方への拡張を試みるが，手掛かり句タイプの候補〈8〉に到達する．手掛かり句タイプの方が優先順位が高いので，文〈9〉の候補は棄却される．

最後に，この段階まで棄却されていない候補を大局話題の導入部として認定する．図1のテキストでは，文〈1〉,〈8〉,〈10〉,〈11〉,〈12〉が認定される．

### 3.2 大局話題の決定

まず，各話題導入部に含まれる顕著名詞句に対して，大局話題としてのもっともらしさを示す大局話題コストを暫定的に割り当てる．この値は小さいほど，話題としてもっともらしいものとする．「は」以外の明示マーカによって提示されているか，またはタイトルに同じ名詞句が含まれかつその名詞句は他の名詞句の一部になっていないか，または固有名詞を含むような顕著名詞句に対しては，コスト値として1を割り当てる．それ以外に対しては2を割り当てる．

図1の文〈1〉の話題導入部では，単文〈1-1〉の顕著名詞句「電話」,〈1-2〉の「サービス」,〈1-3〉の「PCなど」,〈1-4〉の「データ通信」のすべてに対して暫定値2が割り当てられる．ただし，顕著名詞句「電話」と「サービス」はタイトル中の名詞句「電話サービス」に含まれるが，その一部になっているので，話題コストは2となる．他の話題導入部に含まれる顕著名詞句に対しても同様に暫定値が割り当てられる．

次に，連体修飾関係に基づいてコスト値の修正を行う．他の単文によって連体修飾されている顕著名詞句は，説明が加えられた重要な語句であると考えられる．この考えの真偽も含めて，以下の方法を4.2節で検討する．

(方法1) コスト修正を行わない．

(方法2) 他の顕著名詞句を連体修飾している単文に含まれる顕著名詞句のコストを0.5増加させる．

(方法3) 他の単文によって連体修飾されている顕著名詞句のコストを0.5減少させる．

方法3を選んだと仮定すると，図1のテキストでは，単文〈1-4〉の「データ通信」と〈1-2〉の「サービス」のコストを1.5に修正する．

表3 手がかり句による大局話題レベル増減規則
Table 3 A global topic level rule by cue phrase.

| 直前の話題導入部 ＼ 今回の話題導入部 | 入れ子開始、箇条書き項目（第1項のみ）箇条書き全体 | 話題転換 | 入れ子終了 |
|---|---|---|---|
| リード、章立て、箇条書き項目、入れ子開始 | +1 | 0 | 0 |
| 話題転換 | +1 | 0 | 0 |
| 入れ子終了 | +1 | -1 | -1 |

最後に，話題を決定する．各導入部において，コスト最小の顕著名詞句を話題とする．もし，コスト最小のものが複数ある時は，全体タイプの導入部の場合は時間的に最も遅く出現したものを，それ以外のタイプの場合は最も早く出現したものを選択する．

図1の文〈1〉の話題導入部では，単文〈1-4〉の「データ通信」と〈1-2〉の「サービス」のコストが1.5で最小である．両者を比較すると，「データ通信」の方が先に出現しているので，話題として選ばれる．他の導入部については，〈8-1〉の「モバイル・コンピューティング」,〈10-2〉の「ISDN公衆電話」,〈11-1〉の「携帯電話」,〈12-2〉の「簡易型携帯電話」が選ばれる．

### 3.3 大局話題スコープの決定

まず，リードと章立ての両タイプの話題のレベルを求める．リード・タイプの話題が存在すれば，レベルとして1を割り当て，もし存在しなければ，最初の章立てタイプの話題のレベルを1とする．それ以降の章立て話題のレベルは，章番号の遷移に従う．たとえば，「第1章」から「1.1節」へ遷移したら，1増やす．

次に，手掛かり句，箇条書きの全体，項目の各タイプの話題に対して，表3の規則によってレベルを決める．もし今回の話題導入部が全体タイプか入れ子開始型の手掛かり句タイプであるか，箇条書き内で1番目の項目タイプであれば，直前の話題のレベルに1を加えた値を現在のレベルとする．さらに，箇条書き内の2番目以降の項目タイプにも同じレベルを割り当てる．今回が話題転換型か入れ子終了型の手掛かり句である場合に，直前が入れ子終了型でなければ今回は直前と同じレベルとし，入れ子終了型であれば，直前から1引いた値を今回のレベルとする．

図1のテキストでは，文〈1〉の「データ通信」はリード・タイプであるので，レベルを1とする．〈8〉の「モバイル・コンピューティング」は話題転換型の手掛かり句タイプであるので，レベルは1となる．また，

〈10〉の「ISDN公衆電話」,〈11〉の「携帯電話」,〈12〉の「簡易型携帯電話」のレベルはすべて2となる.

最後に,レベルに基づいて話題スコープを決定する.スコープは話題導入部の開始点から始まる.リード,章立て,手掛かり句の各タイプの話題については,自分のレベル以下の話題が現れるまでがスコープである.全体タイプは箇条書きの終わりまでである.項目タイプについては,最後の項目以外は次の項目までであり,最後の項目には箇条書きの終わりまでである.

図1では,「データ通信」のスコープは〈1〉から〈7〉まで,「モバイル・コンピューティング」は〈8〉から〈13〉まで,「ISDN公衆電話」は〈10〉,「携帯電話」は〈11〉,「簡易型携帯電話」は〈12〉と〈13〉になる.

### 3.4 局所話題の抽出

まず,局所話題の導入部候補を抽出する.パラメータ *para-size* と *intro-size* の値が与えられているとして,*para-size* 以上の単文数を含み,かつ大局話題導入部を含まない段落の最初の *intro-size* 個の単文を候補とする.大局話題では明示的手掛かりのある1文を導入部としたが,局所話題ではすぐに話題が提示されるとは限らないので単文数で導入部の大きさを定めた.

パラメータ値を $para\text{-}size = 5$, $intro\text{-}size = 4$ と仮定すると,図1のテキストでは,段落2の単文〈3-1〉から〈4-1〉までと,段落3の〈5-1〉から〈6-2〉までが選ばれる.

次に,話題導入部の各候補に含まれる顕著名詞句に対して,局所話題としてのコストを暫定的に割り当てる.疑問表現をともなう顕著名詞句のコストは1とする.現在,「尋ねる」,「問う」,「聞く」,「質問する」,「質問」が疑問表現としてシステムに登録されている.「は」以外の明示マーカで提示されているか,またはタイトルに含まれているか,または固有名詞を含むか,または直前の大局話題導入部に含まれているもののコストは2とする.それ以外は3である.この割当て後,大局話題の場合と同様に,連体修飾に基づいてコスト値の修正を行う.

コスト修正方法3が選ばれたと仮定すると,図1のテキストの最初の導入部候補では,〈3-1〉の「電話」のコストは3,〈3-2〉の「パソコン通信やインターネットの利用者」は2.5,〈3-3〉の固有名詞「サービスA」は1.5となる.2番目の候補では,〈5-1〉の「サービスB」は2,〈5-2〉の「データ通信の妨げ」と〈6-1〉の「話中」は3,〈6-2〉の「別の電話」は2.5である.

次に,局所話題コストの値が2以下である顕著名詞句を1つ以上含む候補を,話題導入部として認定する.図1のテキストでは両候補とも認定される.

次に,大局話題と同じ方法で,局所話題を決定する.図1のテキストでは各導入部でコスト最小の,〈3-3〉の「サービスA」と〈5-1〉の「サービスB」が話題となる.

最後に,レベルとスコープを決定する.大局話題は入れ子構造を持っていることから,その中にある局所話題どうしには入れ子関係が少ないと思われる.我々の目的は話題構造を100%の精度で抽出することではなく,テキストの概要把握に役立つ精度で抽出する実用的な方法を確立することであるので,局所話題間の入れ子構造は抽出しないこととする.

したがって,局所話題のレベルはその時点での大局話題のレベルの最大値に1を加えた値とする.局所話題のスコープの開始点は各導入部の開始点であり,終了点は次の大局または局所話題の直前までである.図1のテキストでは「サービスA」のスコープは〈3-1〉から〈4-4〉まで,「サービスB」は〈5-1〉から〈7-4〉までとなる.

### 3.5 大局話題と局所話題の統合

大局と局所の両話題を統合した結果の話題構造に対して重複話題の有無を調べる.ここで重複話題とは同レベルでスコープが隣接しているか,直接の親子関係であるかのいずれかが成り立ち,かつ字面が同じである2つの話題のことである.重複話題が検出されたら,後ろに出現する話題を顕著名詞句から削除してから話題構造抽出処理をやり直し,検出されなかったら,その時点での話題構造を最終的な結果とする.

図2の話題構造は図1のテキストについての統合結果である.重複話題は検出されないので,図2が最終的な話題構造となる.

## 4. 評価実験

### 4.1 実験の概要

本論文で提案した方法に基づくシステムを用いて,その有効性を確認するための実験を行った.抽出した話題構造が妥当かどうかの判断は,人間が抽出した話題構造との比較によって行った.テキストとして用いた新聞記事113件は,694段落,1,444文,86,934文字を含んでいる.これらのうち63件をパラメータ値決定などの訓練用に,残りの50件を評価用に用いた.

実験材料に新聞記事を選んだ理由は,情報伝達が目的であり要因3が成り立つことと,入手が容易であることである.また,章立ては含まれていないが,各記事は章に含まれる文章に相当すると考えられるので,有効性の確認には問題ないと判断した.

話題とスコープの精度として再現率と適合率を用い

た．人間とシステムのそれぞれが抽出した話題構造をHとS，その共通部分をIとすると，I/Hが再現率，I/Sが適合率である．ただし，スコープのIは，話題とスコープの両方が一致している部分である．また，スコープの精度は話題構造のレベル付けに関する精度を反映している．すなわち，すべてのレベル付け構造が正しければスコープの適合率も再現率も100%となるが，誤ったものが含まれていれば，適合率と再現率の少なくとも一方は下がる．

ところで，我々のシステムでは話題構造抽出の前処理として，テキストの単語分割などを行う形態素解析を行っている．形態素解析では，単語を辞書にあらかじめ登録しておき，入力テキストと辞書内の単語を照合することにより，単語分割や品詞同定などを行う[6),7)]．未知語の問題，すなわち，辞書に登録されていない単語がテキストに含まれていると単語の区切りや品詞の同定ができないという問題は，形態素解析誤りの最大の原因であり，辞書との照合を行うという方式上，逃れることができない．さらに，未知語には固有名詞が多く含まれているが，固有名詞という品詞情報は話題コストを決める条件に含まれているので，話題構造抽出の精度に大きな影響を与える可能性がある．

この実験では，まず，形態素解析誤りを含まない理想的状況のもとで，訓練用データを用いたパラメータ値などの決定と，評価用データを用いた有効性の確認を行った．次に，未知語が話題構造抽出精度へ及ぼす影響を調べた．

### 4.2 訓練用データによるパラメータ値の決定

パラメータ *item-size-1*, *item-size-2*, *para-size*, *intro-size* の値と，連体修飾に基づく話題コスト修正方法を決定する．テキストの概要把握という利用目的を考慮すると，1つの方針は，訓練用テキストに対する話題スコープの適合率と再現率の平均が最大になるようにパラメータ値などを決定することである．ここで，スコープだけを用いたのは大きな話題が正解していた方が，人間にとって分かりやすいからである．

しかしながら，人間とシステムがそれぞれ抽出した話題構造を比較するには，単純な字面ではなく，意味的に比べるべきである．したがって，比較に際しては人間が介在する必要があるため，各パラメータの値やコスト修正方法のすべての組合せを調べて，最適な値と方法を探すことは現実的に困難である．

我々は，最初に，訓練データに含まれる箇条書きと人間による話題構造の関係を調べることにより，*item-size-1* と *item-size-2* の値を決定した．訓練用データには全部で8個の箇条書きが含まれており，そのすべてについて対応する箇条書き全体話題と項目話題は存在しなかった．これは，新聞記事では，箇条書きは単語や非常に短い文を列挙するために用いられるので，人間はそこから話題を抽出しないためであると考えられる．このことは，話題構造抽出を行った被験者へのインタビューでも裏付けられている．したがって，*item-size-1* と *item-size-2* の両方の値を100に設定して，箇条書きを手掛かりとしないようにした．この実験で用いた新聞記事では，箇条書きは話題構造の手掛かりとはならなかったが，経験的には，たとえば電子メイルや電子ニュースでは非常に有効な手掛かりである．

表4 評価用テキストに対する精度
Table 4 Recall and precision ratios for test texts.

| | 再現率 | 適合率 |
|---|---|---|
| 話題 | 104話題/204話題 =51.0% | 106話題/191話題 =55.5% |
| スコープ | 2172話題/3657話題 =59.4% | 2172話題/3677話題 =59.1% |

次に *intro-size* と *para-size* の値を組合せ的に変化させたときの，話題導入部の適合率と再現率の和を調べた．*intro-size* の値を2から10まで1刻みに変化させ，それぞれの場合について *para-size* の値を3から7まで1刻みに変化させた．全部で45通りの組合せを調べた結果，*intro-size* = 5, *para-size* = 5 の場合に和が最大になったので，これらの値に決定した．

最後に，3通りのコスト修正方法を適用した．スコープの再現率と適合率の合計は，方法3「他の単文によって連体修飾されている名詞句のコストを0.5下げる」の場合に最大となったので，方法3に決定した．

### 4.3 評価用データに対する精度

まず，評価用データに対する平均精度を表4に示す．話題の再現率と適合率はそれぞれ51.0%と55.5%で，スコープは59.4%と59.1%である．ここで，定義式I/HとI/Sから再現率と適合率の分子の値は同じはずであるが，表4における話題の算出式では異なっている．これは，たとえば，人間が話題T1を抽出し，かつそのスコープ内でシステムがT1，T2，T1のように1つの話題T2を間に挟んで同じ話題T1を抽出した場合に，人間による話題Hから見るとIは1個と数え，システム話題Sから見るとIは2個と数えたからである．

次に，精度の分布について考察する．話題に関して適合率と再現率のいずれかが70%以上である精度の良いテキストは50件のうち18件であり，スコープに関しては21件であった．すなわち，半数弱の場合に

我々の手法によってうまく抽出できている．逆に，話題に関して適合率と再現率の両方が30%以下である精度の悪いテキストは2件だけであり，スコープに関しては4件だけであった．このことから，著しい抽出誤りは少ないことが分かる．

スコープの精度が悪いテキストについて誤りの原因を調べると，話題導入部はいずれのテキストでも60%以上の精度で抽出されており，名詞句からの話題選択の失敗が精度を低下させているということが分かった．話題選択の失敗原因のひとつは，場所や手段を提示する格助詞の「で」である．「米国で開発する」のように「で」が場所を提示することは多いが，場所は固有名詞であるので，システムはこれを話題として抽出しやすい．ところが，それは実際には誤りであることも多いので，現在のシステムでは話題マーカには含めていない．したがって，「で」によって本当の話題が提示されていても，現在のシステムでは抽出できない．これらへの対応は今後の課題である．

最後に，我々の方法の有効性について検討する．システムが抽出した話題構造は誤りも含んでいるが，9割強の場合は悪い精度ではないので，人間の優れた推定能力によって，誤りを検出し正しい話題構造を推定することが可能である．また，1割弱の場合は精度が悪いために，正しい話題構造の推定が困難かもしれない．しかしながら，その場合にも話題導入部の多くは正しく抽出されているので，正しい話題は誤ったものの周辺に提示されていると考えられる．我々はユーザに対して話題の名詞句を単独で表示するのではなく，その前後数文字も併記するKWIC (Key Word In Context) 形式で表示している．経験的には，KWIC中の情報から正しい話題構造を推定可能である．このように，テキストの概要把握と情報選択の支援という目的からすると，我々のシステムが抽出する話題構造はユーザが利用可能な精度である．

我々の方法において精度を向上させるためには，より有効な手掛かりを利用する必要がある．今後，利用可能な汎用的手掛かりとして，インターネットのWorld Wide Webでテキストなどを記述するために用いられているHTML表記が考えられる．HTMLでは，強調したい単語を〈em〉と〈/em〉で囲むことによって明示的に示すことが可能であり[8]，もし，これを手掛かりとして利用できれば，話題選択の精度が向上するだけでなく，精度の悪いテキストの数を減らせることが期待できる．

### 4.4 未知語による影響

我々の形態素解析システムの辞書には172,249単語が登録されている．そのうち名詞は135,791語であり，名詞のうち固有名詞は81,980語である．

表5 評価用テキストに対する精度（未知語を含む場合）
Table 5 Recall and precision ratios for test texts including unknown words.

| | 再現率 | 適合率 |
|---|---|---|
| 話題 | 105話題/204話題 =51.5% (Δ0.5%) | 106話題/188話題 =56.4% (Δ0.9%) |
| スコープ | 2175話題/3657話題 =59.5% (Δ0.1%) | 2175話題/3665話題 =59.5% (Δ0.4%) |

評価用データ50件のうち38件に未知語が含まれていた．1〜2語の未知語を含むデータは25件，6語以上含むものは5件だけであり，大半のデータは少数の未知語しか含んでいなかった．未知語の総数は99語あり，その内訳は固有名詞が67語，一般名詞が32語であった．

これらの未知語を形態素解析用の辞書に登録しない状態での，話題構造抽出の平均精度を表5に示す．話題とスコープに対する再現率と適合率ともに，わずかに良くなっているが，この程度の差は偶然であると考えられる．むしろ，この比較によって，未知語があっても平均精度はほとんど変化しないと見なすべきである．

未知語による影響を個別に見ると，精度が変わったテキストは3件だけであり，そのうち1件は精度が下がったが，2件は上がっていた．これら3件はいずれも未知語を多く含んでおり，精度が下がったものは5語，上がったものは8語と6語の未知語を含んでいた．このことから，未知語が多いテキストは精度に影響を受けやすく，少ないテキストは受けにくいと推測できる．未知語が少ない場合に精度への影響がない理由のひとつは，たとえば人名で姓と名の片方が未知語になっていても，片方が固有名詞として認識されるので，「固有名詞を含む」というコスト条件が正しく満たされるからである．この実験で用いた新聞記事では未知語の数は少なかったが，未知語が非常に多いテキスト群に対しては，品詞推定処理が必要である．あるいはテキストの種類によっては，未知語を固有名詞として扱うという簡単な処理で対応できるかもしれない．

## 5. まとめ

テキストの概要の直感的な把握や取捨選択を支援することを目的とした話題構造の抽出方法を提案した．我々の方法は，文書の論理構造，手掛かり句，疑問表現，話題マーカ，連体修飾関係，品詞情報などの書き言葉に汎用的な言語現象を規則化することにより，多

種多様なテキストから話題構造を抽出する．

現実のテキストを用いた評価実験によって，我々の方法は人間にとって利用可能な精度で話題構造を抽出できるということを示した．さらに，現実のテキストでは避けられない未知語の問題に対して，我々の方法は悪影響を受けにくいという見通しも得た．

我々のシステムはWorld Wide Webや電子メイルのインタフェースとして組み込んで，利用されている．自然言語処理の発展のためには，このように実世界で実際に動作するシステムを構築して，有効性を確認するアプローチも重要であろう．

## 参考文献


1) Miike, S., Itoh, T., Ono, K. and Sumita, K.: A Full-Text Retrieval System with a Dynamic Abstract Generation Function, *Proc. SIGIR94*, pp.152–161 (1994).
2) Harn, U.: On Text Coherence Parsing, *Proc. COLING-92*, pp.25–31 (1992).
3) 山本，増山，内藤：手がかり語を用いた日本語文章の段落分けに関する実証的考察，電子情報通信学会言語理解とコミュニケーション研究会，NLC91-15, pp.65–72 (1991).
4) Frohlich, D. and Luff, P.: Applying the Technology of Conversation to the Technology for Conversation, *Computer and Conversation*, Luff, P., Gilbert, N. and Frohlich, D. (Eds.), pp.187–220, Academic Press (1990).
5) Walker, M., Iida, M. and Cote, S.: Centering in Japanese Discourse, *COLING-90* (1990).
6) 吉村，日高，吉田：文節数最小法を用いたべた書き日本語文の形態素解析，情報処理学会論文誌，Vol.24, No.1, pp.40–46 (1983).
7) Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, *Proc. COLING-94*, pp.201–207 (1994).
8) 吉村，家永，鎧：インターネットホームページデザイン，翔泳社 (1995).


(平成8年3月5日受付)
(平成8年9月12日採録)

**竹下　敦**（正会員）

1963年生．1986年京都大学工学部情報工学科卒業．1988年同大学院修士課程修了．同年日本電信電話株式会社に入社．以来，1996年2月までNTTヒューマンインタフェース研究所で自然言語処理や言語コミュニケーション，知的マルチメディア情報検索の研究に従事．現在，NTT北海道法人営業本部マルチメディア推進部に勤務．ACM会員．

**井上　孝史**（正会員）

1965年生．1990年京都大学工学部電気系学科卒業．1992年同大学院修士課程修了．同年，日本電信電話株式会社に入入社．現在，NTTヒューマンインタフェース研究所に勤務．情報検索，自然言語処理などの研究に従事．

**田中　一男**（正会員）

1957年生．1979年神戸大学工学部電子工学科卒業．1981年同大学院修士課程修了．同年，日本電信電話公社に入社．1988～1990年スタンフォード大学客員研究員．現在，NTTヒューマンインタフェース研究所主幹研究員．情報検索などの情報資源活用技術の研究に従事．人工知能学会，日本認知科学会，AAAI各会員．