モノローグに対するブラウジング支援のための話題構造抽出

竹下 敦[†] 井上孝史[††] 田中一男[††]

モノローグ・データに対する，ロバストで適用領域の広い話題構造抽出手法と，その話題構造を利用したブラウジング支援方法「速覧」を提案した．この話題構造抽出手法は，話し手は話題構造を聞き手にうまく伝達するために，言語的手掛かりを意図的に用いたり，あるいは結果的に言語的振る舞いが生じるはずであるという仮定に基づいており，観測可能なそれらの言語現象を体系付けて規則化することにより抽出を行う．たとえば，我々は，「まず」「次に」などの手掛かり句，「これは」などの話題継続句，「について」「が」などの話題マーカ，文などの長さなどの言語現象に着目した．この話題構造抽出は，3ステップから構成される．第1ステップでは手掛かり句などの明示的手掛かりをともなって展開される大局話題を抽出する．第2ステップでは各大局話題内で，明示的手掛かりなしで展開される局所話題を抽出する．第3ステップでは大局話題と局所話題を統合して全体の話題構造を得る．人間が抽出した話題構造とこの方法により抽出したものとを比較する評価実験を行い，この方法の有効性を示した．実際の言語データについて，話題スコープの再現率と適合率は0.585と0.618であった．これは，人間が利用が可能な精度である．

Recognizing Monologue Topic Structures for a Browsing Interface

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

This paper introduces a new tool for browsing called "skim viewer" which uses a new method for monologue topic structure extraction. The extraction method is based on the hypothesis that linguistic clues are used for, and linguistic behavior results from, communicating topic structures. To develop recognition rules, the method incorporates linguistic phenomena such as cue phrases like "mazu (first)" and "tsugi ni (next)", topic continuous expressions like "kore wa (this is)", topic markers like "ni tsuite" and "ga", and sentence length. The method is practical and applicable to a wide variety of monologue data. The method consists of three steps. The first step extracts global topics from explicit clues such as cue phrases. The second step extracts local topics without such explicit clues in each global topics. The third step combines the global and the local topics into a complete topic structure. The effectiveness of the method is shown by comparing manual and system topic structures. Recall and presicion ratios for topic scopes are 0.585 and 0.618.

1. はじめに

電子化された言語情報の流通量が増大するにつれて，情報過多という問題が発生しており，人間は本当に必要な情報を探すために多大な労力と時間を費やしている．情報過多を解決するため，従来から検索，フィルタリング等の研究が行われてきたが，十分な成果が得られていない．たとえば，ほとんどの検索結果には無関係なものが含まれいるので，ユーザは検索結果テキストを読んで，本当に必要な情報だけを選択するというユーザ主導の選択を行わなければならない．

このユーザ主導の選択は時間と労力を要するが，特に，話し言葉を文字起こしした議事録や講演録などのテキストに対しては，読み手の負担はさらに大きくなる．その最大の原因は，話し言葉テキストには，書き言葉のような章立てや段落といった論理構造が与えられていない場合もあり，与えられていたとしても非常におおざっぱであったり，いい加減である場合が多いことである．また，講演録などが非常に長いということも読み手の負担を大きくする原因のひとつである．

本論文では，モノローグにおける話題構造を自動抽出するロバストで適用範囲の広い方法と，ユーザ主導

† 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

の選択における労力を軽減するために，話題構造を章立て構造や目次としてユーザに提示するブラウジング支援方法を提案する．まず話題構造とブラウジング支援方法について述べ，次に，モノローグに対する話題構造抽出の処理について説明し，最後に評価実験の結果について考察を行う．

2. 話題構造を用いたブラウジング支援

2.1 言語情報の基盤としての話題構造

我々の目的は様々な内容の言語データに対するブラウジングを支援することにあるので，大部分の言語データに存在し，かつその内容を読み手に対して直感的に伝達できるような文脈情報を抽出する必要がある．これまで自然言語理解や合成の研究分野で提案されてきた意図構造[1]，議論構造[2]，修辞関係理論[3]などの文脈構造はこの目的に適さない．F. Danes によって提案された3つの主題進行パターン[4]は言語データの内容を反映しているという点では我々の目的に合致するが，主題進行パターンのひとつでは，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かなり深い理解を行う必要があるので，多種多様な言語データに適用できる手法を構築することが困難である．

我々は実際の言語表現として表れている構造で，主題進行パターンを反映したものをターゲットとした．すなわち，人間に言語データを与えて，「同じことが書いてあるブロックと，データ中の名詞句で，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，これを用いることにした．

図1のモノローグ例に対する話題構造の例を図2に示す．この例では，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おり，前の数字はモノローグ全体における文番号を，後ろは各文における単文番号を示すものとする．簡単のために，この例では単純な文のみを用い，「サービスAが（新規通信業者に対抗するために）開始された」のように，単文が別の単文の中に埋め込まれるような文は含んでいない．また，一部を“**”で省略した．

話題構造は，何に関するものかを示す「話題」とそれ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話題の入れ子関係が生じるが，これは「話題レベル」と呼ぶ．一番外側の話題の話題レベルを1とし，入れ子の内側ほど値が1ずつ増加するものとする．たとえば，図2の話題構造において，話題「通信サービス」のスコープは単文 (1-1) の始まりから (7-5) の終わりまでであり，話題レベルは1である．話題「色々な新規サービス」は単文 (2-1) の始まりから (4-2) の終わりまでのスコープで，レベルは2である．

通信サービスの動向
(1-1) 通信サービスは様変わりしてきてますが、(1-2) **。
(2-1) まず、色々な新規サービスがありまして、(2-2) 幅広く利用されています。
(3-1)新規通信業者に対抗するために、(3-2) サービスAが昨年開始されました。
(4-1) これは**、(4-2) **。
(5-1) 次に、従来からのサービスについてですが、(5-2) **。
(6-1) これは**、(6-2) **、(6-3) **。
(7-1) サービスBは10年前に始まりましたが、(7-2) 今では利用者数が減り、(7-3) コストも**、(7-4) ** (7-5) **。

図1 モノローグ例
Fig. 1 An example of a monologue.

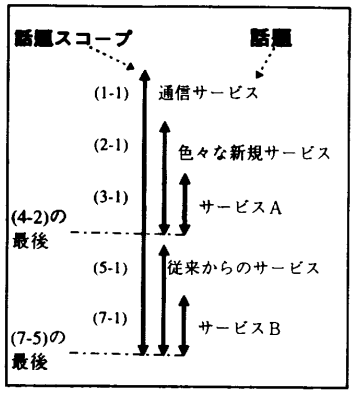

図2 人間による話題構造抽出の例
Fig. 2 An example of manual topic structure.

人間の言語コミュニケーションにおいて，入れ子の外側の話題は内側の話題の意味を限定する．同様に，各話題は，その中の本文の語を限定する．たとえば，「ポスト」という語は，「郵便」の話題の中で用いられた場合は「郵便ポスト」を指すが，「企業内のリストラ」の話題の中で用いられた場合は「地位」を指す．多義語に限らず，一般の言葉の意味も話題構造によって明確になる．意味の限定という作用は抽出した話題構造においてだけではなく，人間の言語コミュニケーションにおいても重要な役割を果たすものと考えられる．

2.2 速覧によるブラウジング支援

言語データから抽出した話題構造を提示することにより，ユーザがその概要を把握することを支援できるが，我々はこれを「速覧」と名付けた．速覧インタフェースとして，図2のように話題構造を目次形式に

した速覧目次や，各話題を章タイトルとして言語データ中に埋め込んだ速覧テキストや，速覧目次の各話題から速覧テキストの対応部分にリンクを張ったハイパーテキスト形式が考えられる．

概要をユーザに提示するという意味での類似技術に，要約自動作成がある[5]．両者の違いは，速覧が名詞句である話題とその構造だけを提示して直感的な概要把握を支援するのに対して，要約は何がどうしたという文章を生成・提示して概要把握を支援するという点である．

したがって速覧は要約よりも情報量が少ないが，興味のある話題を見つけたら，ユーザは言語データ中でその話題に対応する部分だけを読んで，その話題について詳しく知ることができるという特長がある．しかも，話題構造によって前後の内容を直感的に把握できるので一部分だけ読んでもある程度の意味は分かる．

3. モノローグにおける話題構造抽出へのアプローチ

3.1 話題構造抽出の着眼点と処理の流れ

話題構造抽出に関連深い従来技術に，文脈構造や2.1節で述べた主題進行パターンの抽出があり，前者には黒橋らによるロバストな方法[6]があり，後者にはU. Harnの方法[7]がある．また，言語データにおける意味的な区切れ目を認識するセグメント分割も関連深いが，代表的なものとして，J. Morrisらの方法[8]やM.A. Hearstらの方法[9),10]がある．

これらの方法では，文の意味的なまとまりである結束性[11]を計算することが処理の中心である．このため，M.A. Hearstらの方法以外では，言語データの内容の分野知識やシソーラス，類語辞書などを必要とするため，非常に限定された分野に関する言語データしか扱えないという問題がある．M.A. Hearstらの方法は様々な分野の言語データを扱うことが可能であるが，単にセグメント分割するだけであり，我々のターゲットである話題構造のような入れ子のある構造に適用することはできない．また，上記の全方法において，結束性を求めるための処理量が多いという問題がある．

我々は多様な内容の言語データを扱える現実的なシステムを構築するために，結束性ではなく話題導入の際の言語現象に着目した手法を提案する．2.1節で述べたように，話題構造は言語コミュニケーションにおいて重要な役割を果たす．したがって，話題構造を相手にうまく伝達するために，話し手は「次に」のような言語的手掛かりを意図的に用いたり，結果的に言語的振る舞いが生じるはずである．我々は，話題構造抽出を行うために，これらの言語現象を規則化した．

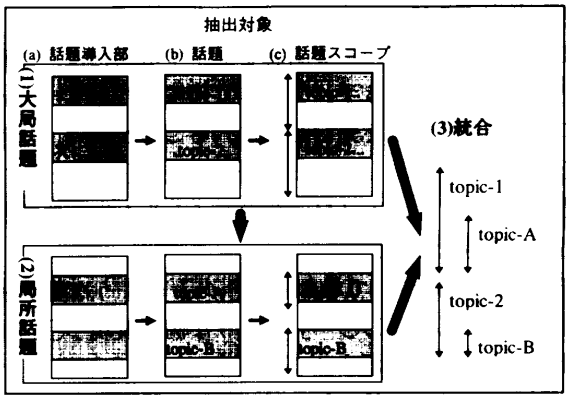

図3 話題構造抽出の流れ
Fig. 3 Outline of topic structure extraction.

話題構造抽出の流れを図3に示す．大局的な話題構造を伝達するためには，「まず」「次に」のような明示的手掛かりが用いられるが，このような話題を「大局話題」と呼ぶことにする．ところが，すべての話題が手掛かりによって明示的に示されるわけではない．これは，手掛かり句などの入れ子が深くなりすぎると，その対応を取るのが困難になり，かえって聞き手や読み手の負担になるからと考えられる．大局話題の中では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．

まず，前処理として形態素解析を行い，モノローグを単語に分割し，各単語の品詞等を同定する．次に，抽出処理としては，大局話題の抽出，局所話題の抽出，両者の統合を順次行う．最初の2つの各抽出処理においては，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．

3.2 モノローグでの話題展開にともなう言語現象

まず，モノローグ中に出現する言語現象のうちで，我々が話題展開の手掛かりとして着目したものについて説明する．

(1) 「まず」「次に」などの手掛かり句：話題の導入だけでなく，入れ子関係も明示的に示す．

(2) 「これは」「この結果」などの話題継続句：前の文から話題が継続していることを明示的に示す．

(3) 「について」や「が」「を」などの話題マーカ：マーカによって，話題を提示しやすさに優先順位がある．

(4) 文などの長さ：新しい話題を提示する場合は，聞き手にその話題を確実に伝達しようとするために，文が長くなると考えられる．1文の長さだけでなく，話題継続句で結合された文の集合である疑似段落の長

さも考慮する．

(5) 先行要約：大きな話題を提示する際には，これから話す内容を聞き手により良く理解してもらうために，そのあらましを先行して述べることが多い．この先行要約中に出現している単語は，後で子話題として提示されやすい．モノローグのタイトルも先行要約と同様の働きをする．

(6) 「○○について尋ねると，…」のような疑問表現や「たとえば，○○…」のような例示表現：○○に関する子話題を提示することが多い．

次に，モノローグにおいて，上記の言語現象を生じさせる要因をあげる．

- **要因 1 記録性のなさ**
- **要因 2 インタラクションのなさ**
- **要因 3 話題を伝えたいという意志**

要因 1 は，原稿や文字起こしテキストであっても，元は話し言葉であることに起因する．要因 2 によって話題展開権を保持できるので，話者は自由に話題を展開できる．要因 3 はモノローグだけでなく，情報伝達を行う言語データ一般における大きな要因である．

最後に，要因と言語現象の対応付けを行う．7 つすべての言語現象は要因 3 の影響を受けている．また，言語現象のうち，明示的に話題展開を示すものは手掛かり句と話題継続句だけであるが，他に明示的な手掛かりがないのは，要因 1 による制限である．たとえば，記録性のあるテキストでは章立てや箇条書き等などの明示的な手掛かりを用いることが可能である．

文などが長くなるという現象は，要因 2 に関連が深いと思われる．すなわち，この現象の原因のひとつは，インタラクションがないために聞き手の理解度が分からず，説明が長くなるのではないかと考えられる．先行要約が可能となったのも要因 2 による．インタラクションがある対話では先行要約は使えない．

これらの対応付けは，我々の話題抽出手法の改良を行う際の指針となる．たとえば，若干のインタラクションを含むが，対話というよりはモノローグに近いような言語データを扱う場合は，要因 2 に対応する文などの長さや先行要約を用いた処理を改良すればよい．

3.3 大局話題の導入部の抽出

第 1 に，話題導入部の候補を検出する．モノローグ開始時には必ず話題が導入されているはずである．また，新しい大局話題の始まりを明示的に示すのは手掛かり句だけである．

表 1 に手掛かり句の一覧を種類別に示す．新たに話題の列挙が開始するとき，最初の話題には「まず」などの入れ子開始型が，最後の話題には「最後に」などの入れ子終了型が，それ以外には「次に」などの話題転換型が用いられる．また，「これは」，「この結果」等の話題継続句は新たな大局話題がないことを示す．現在，登録されている話題継続句は，上記 2 つと，「によりますと」「これに対し」「このため」である．

表 1 手掛かり句
Table 1 Cue phrases.

種類	手掛かり句の例
入れ子開始型	まず，最初に，第 1 に，1 番目，まず第一に
話題転換型	次に，次は，ところで，それから，さて，それでは，さっそく，このあと，そして，あと，もう一つ，続く，今度は，第 2 に，……，第 9 に，2 番目，……，9 番目
入れ子終了型	最後に，終わりに

以上から，(a) 手掛かり句を含み，かつ第 1 単文に話題継続句を含まない文と，(b) モノローグの第 1 文を話題導入部の候補として検出する．

図 1 のモノローグ例において，たとえば単文 (1-1) と (1-2) から成る文を文 (1) と記述すると，第 1 文の文 (1) と，手掛かり句「まず」と「次に」をそれぞれ含む文 (2) と (5) が話題導入部の候補として検出される．

第 2 に，述語を 1 つだけ持つ単位である単文について，そこに含まれる名詞句のうち，単独では意味を表さない代名詞のような語でなく，かつ単文中で最も強調されているものを顕著名詞句として抽出する．名詞句としては常に最長のものを採用するものとする．

強調の度合いは，名詞句を提示するマーカの優先順位によって決定する：

明示マーカ > 非明示マーカ

ここで，明示マーカとは，「について」「は」のように話題を提示するための表現であり，非明示マーカとは格助詞の「が」や「を」のように主語や目的語などの表層格を示す表現である．現在，41 個の明示マーカと 14 個の非明示マーカが登録されている．

非明示マーカ間にも主語，直接目的語，間接目的語という順で優先順位付けが行われている．M. Walker らは，代名詞等の解析を行う局所的な文脈処理のセンタリングにおいて，話者が事象をどこから記述しているかという視点 EMPATHY もマーカ優先順位に組み込んでいる[12]が，我々は EMPATHY は代名詞解釈には有効ではあるが，より大局的な文脈である話題構造には適さないと判断して，これを除いた．

図 1 のモノローグ例において，単文 (1-1) からは明示マーカ「は」によって提示された「通信サービス」

が，(2-1) からは非明示マーカ「が」によって提示された「色々な新規サービス」が顕著名詞句として抽出される．同様に，(3-1) の「新規通信業者」，(3-2) の「サービスA」，(5-1) の「従来からのサービス」，(7-1) の「サービスB」，(7-2) の「利用者数」が顕著名詞句として抽出される．(2-2) からは顕著名詞句は抽出されない．

第3に，話題導入部の候補に含まれる顕著名詞句に対して，大局話題コストを求める．コストは小さいほど大局話題の話題として尤もらしい．もし，顕著名詞句が明示マーカで提示されているか，固有名詞を含むか，モノローグのタイトルに含まれるかのいずれかが成り立てば，その顕著名詞句のコストを1とする．固有名詞は指示対象が決まっており，文脈を限定する力が強いので，話題として用いられやすいし，聞き手も話題として認識しやすいからである．また，タイトルは先行要約の働きをするからである．これら以外の顕著名詞句のコストは2とする．

図1の単文 (1-1) の顕著名詞句「通信サービス」のコストは1で，(2-1) の「色々なサービス」は2，(5-1) の「従来からのサービス」は1である．

最後に，話題提示型の同定と話題導入部の決定を行う．モノローグでは要因1「記録性のなさ」と要因2「インタラクションのなさ」の両方が成り立つので，大局話題のすぐ後ろにより詳細な話題を提示する「逐次型」と，複雑な名詞句を使って一度に話題を提示する「一括型」の両方が用いられる．たとえば，「国連の安全保障理事会は，開票作業が進められているカンボジアの総選挙は自由公正に…」という文から，人間は「国連の安全保障理事会」と「カンボジアの総選挙」の2つを話題として抽出したが，逐次型はこのように1つの文に複数の話題が提示される場合である．

もし，1つの話題導入部候補にコスト1の顕著名詞句が複数存在すれば，逐次型とし，それ以外の場合は一括型であるとする．もし，逐次型であれば，その導入部候補中でコスト1の最初の顕著名詞句を含む単文の最後までを話題導入部とする．もし，一括型であれば，候補をそのまま認定する．ここで求められた話題導入部は，その先に述べることを説明する先行要約の機能も果たし，先行要約中に含まれる名詞句は，後で局所話題として選ばれやすくなる．

図1のモノローグ例において，単文 (1-2) からコスト1の顕著名詞句が抽出されないと仮定すると，文 (1) にはコスト1の顕著名詞句が1つしかないので，一括型の話題提示となり，文 (1) が話題導入部として認定される．同様に，文 (2) と (5) の話題導入部候補も一括型であり，導入部としてそのまま認定される．

表2 大局話題でのレベル付け規則
Table 2 A topic level rule for global topics.

前回の手掛かり句 \ 今回の手掛かり句	入れ子開始	話題転換	入れ子終了
入れ子開始	+1	0	0
話題転換	+1	0	0
入れ子終了	+1	-1	-1

3.4 大局話題の抽出と話題スコープの決定

各話題導入部に含まれる顕著名詞句で，大局話題コストが最小のものを話題として抽出する．もし，コスト最小のものが複数ある場合は，要因3によりできるだけ早く話題を提示するはずであるから，最初に出現しているものを話題とする．

図1の例では，文 (1) の話題導入部から「通信サービス」，文 (2) からは「色々な新規サービス」，文 (5) からは「従来からのサービス」が話題として抽出される．

モノローグ開始時の話題の話題レベルは1とする．それ以降の話題については，今回と前回の手掛かり句のそれぞれが，3.3節で説明したどの種類に属するかによって表2の規則を用いてレベルを増減する．今回の手掛かり句が入れ子開始型の場合は話題レベルを1増やす．前回の手掛かり句が入れ子終了型で，今回が話題転換型か入れ子終了型であれば，話題レベルを1減らす．それ以外の場合は話題レベルは変更しない．なお，モノローグ開始時は入れ子開始型として扱う．また，話題スコープは自分以下のレベルを持つ話題の前までとする．

図1のモノローグ例では，話題レベルは話題「通信サービス」が1，「色々な新規サービス」と「従来からのサービス」が2である．また，話題スコープはそれぞれ，文 (1) から (7) の終わりまで，文 (2) から (4) の終わりまで，文 (5) から (7) の終わりまでである．

3.5 局所話題の導入部の抽出

局所話題の導入部の手掛かりとして文などの長さに着目した．まず，以下の条件をすべて満たす文Sを話題導入部の候補として抽出する．ただし，文Sにおいて，逐次型の大局話題が抽出されている場合には，S全体の代わりに，その大局話題を含む単文より後の部分を用いる．

(a) あらかじめ設定された値 *sent-size* 以上の数の単文を含んでいる．

(b) 一括型の大局話題が抽出されていない．

(c) 話題継続句を第1単文に含まない文は新しいブロックとし，含む文はその直前の文のブロックにマージするという規則で求めたブロックについて，文Sはブ

ロックの先頭の文であり，かつブロックはあらかじめ設定された値 *block-size* 以上の数の単文を含んでいる．

次に，候補中の各顕著名詞句に対して局所話題としてのコストを計算する．「尋ねる」のような疑問表現や，「例えば」のような例示表現と同じ単文に含まれている顕著名詞句のコストは1とする．ここで，疑問表現としては「尋ねる」「問う」「聞く」「質問する」の4語が，例示表現としては「例えば」「一つの」「一つには」「一つに」の4語が登録されている．また，「は」以外の明示マーカによって提示されているか，直前の先行要約，すなわち大局話題の導入部に含まれているか，固有名詞を含む顕著名詞句のコストを2とする．上記のいずれの条件も満たさないもののコストは3とする．

最後に，各候補に2以上の話題コストを持つ顕著名詞句が1つ以上あれば，話題導入部として認定する．このように，逐次型の話題提示と認定されている場合でも，必ずしも後続する詳細な話題が局所話題で抽出されているわけではない．

図1のモノローグ例では，*sent-size* = 2，*block-size* = 4 と仮定すると文 (3) と (7) という2つの話題導入部候補が抽出される．単文 (3-2) の「サービスA」と (7-1) の「サービスB」は固有名詞であるので局所話題コストは2となる．また，(3-1) の「新規通信業者」と (7-2) の「利用者数」と (7-3) の「コスト」の話題コストは3となる．両候補ともコスト値が2以上の顕著名詞句を含むので，話題導入部として認定される．

3.6 局所話題の抽出と話題スコープの決定

局所話題における話題は，局所話題コストを用いるという点以外は，大局話題と同様に抽出する．図1の文 (3) の話題導入部から「サービスA」が，文 (7) からは「サービスB」が話題として抽出される．

モノローグでは話者は話題展開権を保持でき，細かな話題は出にくいので，隣接する局所話題は同レベルであるとする．すなわち，局所話題における話題レベルは，その時点での大局話題のレベルの最大値に1を加えた値である．話題スコープは話題導入部の始まりから，それ以降に最初に現れる局所または大局話題までである．

図1のモノローグ例では，話題「サービスA」と「サービスB」のレベルはともに3であり，スコープはそれぞれ文 (3) の始まりから (4) の終わりまでと (7) の始まりから (7) の終わりまでである．

3.7 大局話題と局所話題の統合と再計算

大局話題と局所話題を統合して，全体の話題構造を完成させ，さらに，全体の話題構造に対して，話題の重複の有無を調べる．話題構造に含まれる任意の2つの話題A，Bについて，Aが先に出現していると仮定する．両者の字面が同一であり，かつ次の条件が成り立てば，AとBは重複していると呼ぶ：(a) AがBの親話題であるか，(b) あるいはAの直後の話題がBであり両者が同レベルである．

人間は重複を含むような話題構造は抽出しないし，速覧の際に重複はユーザの混乱を招く．もし，重複が検出されたら，後に出現するBを顕著名詞句からはずして，話題抽出処理をやり直す．

図1のモノローグ例に対する統合後の話題構造は図2のとおりになる．この話題構造からは重複は検出されないので，これが最終的な話題構造となる．

4. 話題構造抽出の評価実験

4.1 評価実験の概要

本論文の手法を用いた話題構造抽出システムを構築し，評価実験を行った．実験に使用したモノローグ・データはニュース原稿127件であり，63件をパラメータ調整用に，残りの64件を評価用に用いた．1件の平均文数は8.92で，文字数は884であった．

評価精度としては，あらかじめ人間が抽出した話題構造Mとシステムによる話題構造Sを比較して両者の一致部分Iを求め，人間による話題のうちどれだけがシステムによって抽出されたかを示す再現率I/Mと，システムによる話題のうちどれだけが人間によって抽出されたかを示す適合率I/Sを計算した．評価対象は話題とスコープである．

人間による話題構造Mは，本論文の処理内容を知らない2人の被験者A，Bによって作成された．第1ステップで，被験者Aがモノローグから話題構造を抽出し，第2ステップで，その話題構造を被験者Bがチェックし，自分の意見との不一致部分を抽出し，第3ステップで，それらの不一致部分について被験者AとBが話し合い，最終的な話題構造を決定した．

話題が一致しているかどうかの判定は，最初に計算機で完全に一致する話題を調べ，それ以外のものについて，2人の人間が正しいかどうかを判定し，両者の意見が一致する話題を「一致している」とした．また，話題の一致部分を求める際は，速覧という目的から，図4に示すように，システムが求めた1つの話題に対して，人間による複数の話題が対応する場合や，その逆の場合も一致しているとした．したがって，精度を求める際の一致部分Iが再現率と適合率とで異なる場合もある．図4の例では，再現率では一致話題を3個として数え，適合率では2個として数える．

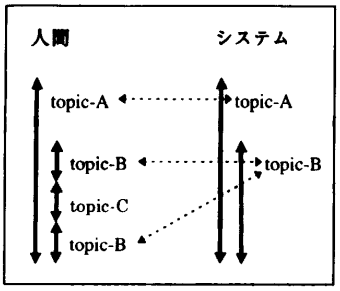

図4 話題の比較における対応付けの例
Fig. 4 Comparison of manual and system topic structures.

スコープの精度を求める際には，スコープの単位には単文数を用い，また，人間とシステムで一致している話題について，スコープの重なりを一致部分として抽出した．これは言い換えると，スコープの長さに応じて話題に重み付けしたものであり，たとえば，長い話題が一致しているほうが良い精度となる．

また，本実験は，我々の方法のみの精度を求めるために，システムで前処理として行っている形態素解析では未知語の問題は起こっていないと仮定する．すなわち，形態素解析ではあらかじめ構築した単語辞書を用いてモノローグ・データの単語への分割，各単語の品詞等の同定を行うので，辞書に登録されていない未知語の品詞を同定は困難であり，大局，局所の両方の話題コストの割当て条件「固有名詞を含んでいる」の判定に影響を受ける可能性がある．しかしながら，本実験では全データについて，あらかじめ固有名詞を抽出，登録しておいた．また，実際のデータには未知語が存在するが，未知語の多くは人名の姓か名前であり，姓か名前のどちらかは単語辞書に含まれている場合が多いので，「固有名詞を含む」という条件は満たされ，未知語の影響は少ないことが予想される．

4.2 パラメータの調整

調整用データを用いて，パラメータ *sent-size* と *block-size* の最適な値を求めた．調整に際しては，速覧という用途を考えて，スコープの再現率と適合率の平均値が最大になるようにした．

sent-size の値を1から6まで，*block-size* の値を2から6まで，それぞれ整数値で変化させ，*block-size* $\geq$ *sent-size* を満たす20通りの値の組合せのすべてについて，適合率，再現率を求めた結果，*sent-size* と *block-size* ともに4という最適値が得られた．この調整結果では *sent-size* の条件が満たされれば，必ず *block-size* の条件も満たされるので，*block-size* の条件は無意味になる．しかしながら，別のデータに対しては値が変わるはずであるし，同じデータでもたとえば話題の適合率が最大になるように調整を行えば，*sent-size* $=4$ と *block-size* $=6$ という値が得られる．また，講演録や議会会議録など異なる種類のモノローグに対しては，たとえば *sent-size* $=6$ と *block-size* $=8$ のように，パラメータ値を大きくするとともに，*block-size* $>$ *sent-size* とした方が，より適切な話題構造が得られるということが経験的に分かっている．

また，数値パラメータの値だけでなく，話題マーカや例示マーカのような辞書項目も決定した．これらは調整用データや，評価実験で用いていないニュース原稿や講演録などから人手で抽出，確認した．1人が抽出したものに対して，他の2人のうち，1人が賛同した表現を辞書に加えた．

表3 評価用データに対する精度
Table 3 Experimental results.

	再現率	適合率
話題	209話題/341話題 =0.613	198話題/344話題 =0.576
スコープ	3596単文/6148単文 =0.585	3596単文/5818単文 =0.618

4.3 評価用データに対する精度

評価用データに対して行った話題構造抽出の精度を**表3**に示す．話題の再現率，適合率はそれぞれ0.613と0.576で，スコープはそれぞれ0.585と0.618であった．速覧という目的を考えると，個々のデータに対する精度の分布も重要である．64件中，話題の再現率と適合率の両方が80%以上のものは10件，60%以上は24件であり，30%未満は3件であった．スコープの再現率と適合率の両方が80%以上のものは5件，60%以上は32件であり，30%未満は5件であった．

この実験結果によると，我々の手法では精度が極端に悪いものは少ないが，精度が非常に良いものも多くはない．多くのデータにおいては，人間が話題として抽出しなかったものを計算機が話題として誤抽出したり，その逆の抽出漏れが何件か起きている．しかしながら，速覧という用途では，誤抽出や抽出漏れがあっても，人間はそのような誤りを見抜き，正しいものを推定できるので，現在の精度でも利用可能である．

誤抽出や抽出漏れの原因を調べた．最大の原因は，手掛かり句や長い文などの言語現象が必ずしも話題構造に対応していないことである．たとえば，「また」という語は話題導入を示すことも多いが，そうでない場合も多い．別の原因として，人間が話題を抽出する際は他の話題との関係で選ぶことがあるということである．たとえば，直前の話題と対比させて新しい話題を

提示する場合は，話題コストがたとえ低くても話題として選ばれる．

最後に，精度を現状よりも向上させる方法について考察する．1つ目の方法は言語現象の精密化である．適用対象を限定すれば，その対象特有の現象を抽出規則に組み込むことが可能である．現システムでも，話題継続句や手掛かり句の辞書や，パラメータを修正することである程度は対処できる．たとえば，議会の議事録を対象とした場合は，議会独特の言い回しを辞書に加えればよい．2つ目の方法は3.1節で説明したセグメント分割等の結束性の解析と組み合わせることである．たとえば話題スコープの誤りの改善が期待できる．

5. まとめ

モノローグ・データの内容の直感的把握を支援する速覧を目的とした，話題構造の自動抽出方法を提案した．我々はモノローグ中に出現する手掛かり句，話題継続句，話題マーカ，文などの長さ，先行要約，疑問表現や例示表現という言語現象に着目し，話題構造の抽出規則を構築した．これにより，広範囲・多種多様なモノローグ・データに対して，話題構造抽出を行い，速覧を提供することが可能となった．

評価実験によって，実際のモノローグ・データについて，話題とスコープの適合率と再現率が57%以上という人間が利用可能な精度で話題構造を抽出できることが確認された．これにより，実用的な速覧システムの可能性が示された．実際的には，適用対象に合わせて話題継続句や手掛かり句の辞書をカスタマイズすることによって，より良い精度を得ることも可能である．また，我々の手法は複雑な処理を含まないので，リアルタイム速覧システムの構築も可能である．

参考文献

1) Grosz, B. and Sidner, C.: Attention, Intention and the Structure of Discourse, *Computational Linguistics*, Vol.12, No.3, pp.175–204 (1986).
2) Cohen, R.: Analyzing the Structure of Argumentative Discourse, *Computational Linguistics*, Vol.13, No.1-2, pp.11–24 (1987).
3) Mann, W. and Thompson, S.: Rhetorical Structure Theory: Description and Construction of Text Structures, *Natural Language Generation*, Kempen, G. (Ed.), pp.85-96, Martinus Nijhoff (1990).
4) Danes, F.: Functional Sentence Perspective and the Organization of the Text, *Papers on Functional Sentence Perspective*, Danes, F. (Ed.), pp.106–128, Academia (1974).
5) Paice, C.D.: Constructing Literature Abstracts by Computer: Techniques and Prospects, *Information Processing and Management*, Vol.26, No.1, pp.171–186 (1990).
6) Kurohashi, S. and Nagao, M.: Automatic Detection of Discourse Structure by Checking Surface Information in Sentences, *COLING-94*, pp.1123–1127 (1994).
7) Harn, U.: On Text Coherence Parsing, *Proc. COLING-92*, pp.25–31 (1992).
8) Morris, J. and Hirst, G.: Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics*, Vol.17, No.1, pp.21–48 (1991).
9) Hearst, M.A. and Plaunt, C.: Subtopic Structuring for Full-length Document Access, *Proc. SIGIR*, pp.59–68 (1993).
10) Hearst, M.A.: Multi-paragraph Segmentation of Expository Text, *Proc. ACL-94*, pp.9–16 (1994).
11) Halliday, M. and Hasan, R.: *Cohesion in English*, Longman (1976).
12) Walker, M., Iida, M. and Cote, S.: Centering in Japanese Discourse, *COLING-90* (1990).

(平成7年11月7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．