自 然 言 語 処 理 83－4
（1991. 5. 16）

表層的処理による話題抽出

竹下 敦
NTTヒューマンインタフェース研究所

談話を話題という側面からとらえた「話題展開モデル」を提案した。このモデルは話題構造を大まかな「分野」、明示的な言語表現で導入される「明示的話題」、明示的話題に関連して展開する「関連話題」、明示的言語表現でによって行われる「明示的話題転換」という要素でとらえる。目的志向対話だけでなく、従来の対話モデルでは扱えないような目的志向性の低い対話も扱うことができる。さらに、この話題展開モデルの談話構造を同定するための「階層型話題同定モデル」という計算枠組みを提案した。また、話題同定モデルによる同定結果と人手による話題同定結果を比較して話題同定モデルの評価を行なった。ある程度の有効性が確認された。

Topic Extraction by Sruface Processing

TAKESHITA, Atsushi
NTT Human Interface Laboratories
1-2356 Take Yokosuka-Shi, Kanagawa 238-03, Japan

A *topic proceeding model* is proposed. The model is composed of *topic fields*, *explicit topics* marked by linguistic expressions introducing topics, *unfolded topics* related to explicit topics and *explicit topic changes*. Not only task-oriented but also less-task-oriented dialogues can be treated. A *hierarchal topic identification model* is also proposed. It is a computational model of the topic proceeding model. From the comparison of topics by the topic identification model and manually identified topics, the model ability is verified.

1 はじめに

リアルタイムに行われる談話状況の一側面の認識として、談話における話題の同定を行う手法について述べる。ここで対象とする談話とは、一人で話すモノローグや、二人以上で相互に言葉を交わす対話であり、話し言葉を用いて行われる。対話は目的志向のものに限らず、解説等の非目的志向のものも対象とする。

従来の談話理解の研究においては、ゼロ代名詞や代名詞の解釈のための極めて局所的な話題焦点の同定や、発話における話者の意図の理解が中心であった。これらの談話理解研究を含めたAI研究においては、問題となっているドメインがある程度分かっていることが前提となっている。しかしながら、大規模なAIシステムではドメインを絞ることができないために推論において膨大な曖昧性が生じ、そのままではAI技術を活用することはできない。これを解決するためには、話題という側面からみた対話の大局的構造の認識を行なうことが必要である。

また、話題という談話の一側面が同定できただけでも例えば以下のようなサービスに役立つ。

1. 話題に応じた情報提供
   対話者の話題を同定し、大規模データベースの中からそれに関連する情報をあらかじめ絞り込んだり、さらには対話者に対して自動的に提供する。情報案内サービスのように、一方の話者からの問い合わせに対して、もう一方の話者が情報検索を行い、その結果を迅速に質問者に返さなければならない場合には特に有効である。

2. マルチメディアへのラベル付け
   各時点での話題を検索用のラベルとして付与する。また、マルチメディア・データを意味のある単位ごとに分割することの支援も可能である。

2 話題面からの談話のモデル化

2.1 話題の定義

直感的には以下のような談話指示対象のうち、そのとき話されているものであり、対話参加者により共有されているものが「話題」である。

1. 名詞句によって表現される事物・事柄
2. 動詞等の述語によって表現される事象

これらの談話指示対象はどれも話題候補となるが、実際に話題として認定されるのは、再び話題として提示されたり、代名詞やゼロ代名詞によって言及されるなどして、しばらくの間継続するものである。

2.2 話題展開モデル

談話において同一の話題が継続している連続した発話の集合を談話セグメントと呼ぶ。談話セグメントは入れ子構造をなしていてもよい。図1の対話例において、談話セグメントS2で話題となっているのは「国内便」、S3では「エアメイル」、S1では「郵便物の出し方」、S4では「切手の購入」である。談話セグメントS1はS2とS3を含む。

談話の進行を、対話参加者によるプランの遂行としてとらえるプラン認識モデルにおいては、談話の進行の仕方の大まかな分類としてプランの継続、転換、詳細化が提案されている [J.Litman & Allen, 1987]。しかしながら、目的志向でない対話ではそもそも対話を明確なプランとしてとらえるのが困難であり、また、とらえることのできる部分に関しても話題の進行はこの分類のどれに当てはまるかが不明確なものが多い。さらに、目的志向型対話に関しても、現実の対話においては例えば、事情や状況を相手に説明するという行為が頻出しており、この説明部分においては目的志向でない対話と同様に話題の進行の仕方は上記の分類には必ずしもあてはまらない。

また、上記の分類ではプランの継続と詳細化を区別しているが、話題に関しては必ずしもisa関係の詳細化や抽象化の方向に進行するわけではなく、むしろある話題が導入されたら、それに関連する事物や事象へと展開する。したがって、以下のような要素から構成される話題展開モデルを提案する。

1. 分野: 話題が属する分野。例えば、野球について話しているのか、ゴルフについて話しているのかという話題に関する大きな分類。
2. 明示的話題: 明示的に導入される話題。例えば、「〜について伺いたい」のような話題を導入する言語表現によって導入される。
3. 関連話題: 明示的話題に関連して展開される話題。
4. 明示的話題転換: 明示的に行われる話題転換。例えば、「まず」「次に」といった言語

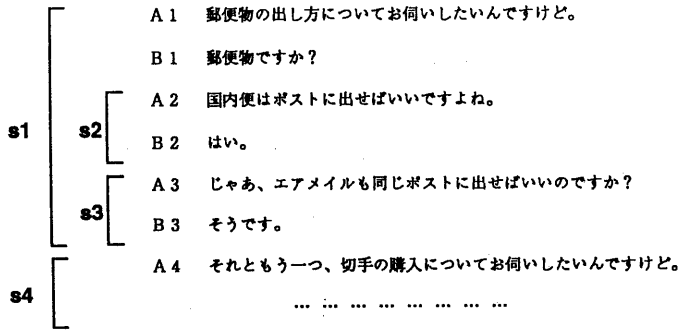

図 1: 対話例

表現によって行われる。

話題展開モデルの各要素は、典型的には図2に示すように展開する。すなわち、明示的話題が導入された後、幾つかの関連話題が展開する。これが何回か繰り返された後に明示的話題展開が行われる。その後、さらに明示的話題・関連話題・明示的話題展開が繰り返され、その後に新しい分野に話題が移って談話が同様に進行する。

図1の対話例においては、話題分野は「郵便業務」、S1 における明示的話題は「郵便物の出し方」、関連話題は S2 では「国内便」、S3 では「エアメイル」であり、また、S4 における明示的話題は「切手の購入」である。このように話題展開モデルの各要素を同定することにより、談話セグメントの構造を同定することができる。

[J.Litman & Allen, 1987] において提案されているプラン認識モデルは3つの要素から構成される: プランの継続・詳細化・転換等を記述する談話プラン、郵便物を出すにはどうすればよいかのように現実世界においてある事柄を遂行するためにはどのようなことをしなければならないかを記述するドメイン・プラン、個々の発話を行うための発話行為。話題展開モデルにおける分野はプラン認識モデルにおけるドメイン・プランの話題的側面からの認識に相当する。また、話題展開モデルにおける明示的話題と明示的話題転換はプラン認識モデルにおける談話プランの「転換」の話題的側面の認識に、関連話題は発話行為の話題的側面の認識に相当する。

3 階層型話題同定モデル

3.1 モデルの概要

「話題展開モデル」の話題構造を同定するために、話題同定処理を3つの階層に分けて行う。図3に示すように、階層1では「野球」や「ゴルフ」といった分野を同定する。階層2では明示的話題と明示的話題転換を同定する。明示的話題転換の間に挟まれた明示的話題は同一の事柄に関するものであり、同一の談話セグメントに属するとみなすことができる。図3では「阪神・巨人戦」「講評」という明示的話題の後に明示的話題転換が行われ、さらにその後「中日・広島戦」という明示的話題が導入される。「阪神・巨人戦」と「講評」は一つの談話セグメントを形成し、「中日・広島戦」は別の談話セグメントを形成する。各談話セグメントの最初の明示的話題は、「阪神・巨人戦」のようにそのセグメントで展開する話題を特徴づけることが多い。階層3では関連話題を同定する。図3では、「阪神・巨人戦」という明示的話題に対して、まず「スコア」という関連話題が展開している。

各階層は、隣接する階層と話題転換に関する情報を交換することにより、話題構造同定の向上させる。

話題同定の処理に対して、リアルタイム性と逐次性という制約を課す。したがって、談話のテキスト全体を調べてから、各時点での話題を同定するというアプローチを採ることはできない。各発話を逐一処理し、その時点での話題を同定する。

また、以下の理由から、ゼロ代名詞や代名詞の解釈処理は行わなかった。

1. フォーマルな対話では、原則的に話題は明示

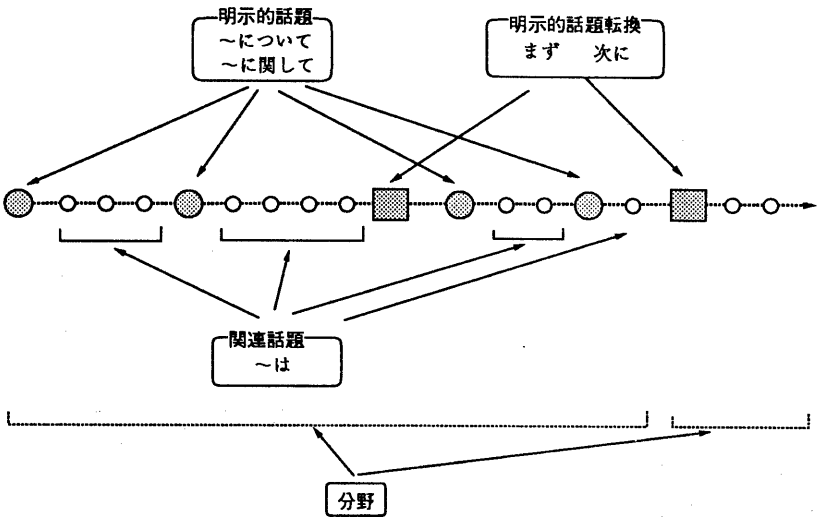

図 2: 話題展開モデル

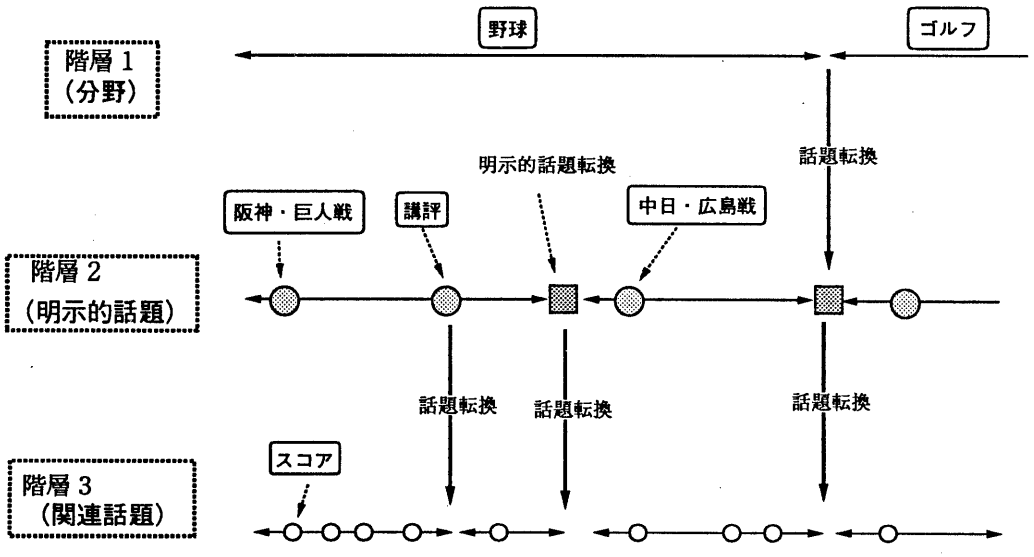

図 3: 階層型話題同定モデルの概要

的に導入されるという仮説を立てた。

2. face-to-face やビジュアル・テレフォンを介したマルチ・モダル対話においては、必ずしも話題が言語表現によって導入されるとは限らない。例えば、画面上の物をいきなり「これ」や「それ」という代名詞で指示することがあり、その後その指示対象が話題となることもある。このような現象は自然言語による発話だけではなく、話者の動作のような画像情報も理解する必要があり現在の技術レベルでは非常に困難である。

3.2 各階層における処理

3.2.1 階層 1: 分野

分野はある特定の分野でのみ通常用いられる分野用語の頻度を調べることにより同定する。例えば、図 4において四角で囲った語が分野用語である。分野用語とそれが属する分野の対応を記述した図 5に示すような対応辞書はあらかじめ用意しておく。

A1 郵便物の出し方についてお伺いしたいんですけど。
B1 郵便物ですか？
A2 国内便はポストに出せばいいですよね。
B2 はい。
A3 じゃあ、エアメイルも同じポストに出せばいいのですか？
B3 そうです。
A4 それともう一つ、切手の購入についてお伺いしたいんですけど。
… … … … … … … …

図 4: 対話例における分野用語

分野用語辞書を用いて形態素解析結果から分野用語を検出し、最も最近検出された n 個の分野用語に対応する分野について最も数の多いものを現在の分野とする。最近の n 個の分野用語を管理するためには長さが n の FIFO(First In First Out) を 1 つ用意する。初期状態では FIFO は空である。FIFO がまだ全部埋まっていない、すなわち FIFO 中のリスト数が FIFO の長さ n より小さいときは分野名は同定できない。なお、n は例えば 5 のような適当な値を設定すればよい

用語	用いられる分野
郵便物	郵便
国内便	郵便、航空
ポスト	郵便、ビジネス
エアメイル	郵便
切手	郵便

図 5: 話題分野の同定に用いる分野用語辞書の例

3.2.2 階層 2: 明示的話題と明示的話題転換

英語では is-a 構文の主語、there 構文の主語、分裂文の分裂要素は話題 (焦点) を明示的に示す [L.Sidner, 1983]。これらに対応する図 6に示すような表現は日本語においても話題を導入する。また、図 7に示すような言語表現も話題を明示的に導入する。

明示的話題の同定においては、まず、図 6と図 7に示すような明示的話題同定用辞書と発話文の形態素解析結果のマッチングをとって、話題候補を検出する。例えば、図 6の「X は Y です。」という文では、X を話題候補として検出する。もし、話題候補が検出され、かつそれが代名詞のような特定の概念を表さない不適切語でなければ、その話題候補を明示的話題の集合に追加する。もし不適切語であるか、あるいは話題候補がなかった場合には、何も行わない。

明示的話題転換の検出は、同様に発話文中に図 8に示すようなクルーワードと呼ばれる語を探すことにより行なう。見つかればそこで明示的話題転換が起きているとみなされる。

主部	述部
(話題) は	～だ
(話題) が、は、も	いる、ある

図 6: 明示的話題の同定用辞書の例 (1)

語句	
(話題) 関して	(話題) について
(話題) とは	(話題) というのは
(話題) といえば	(話題) というと

図 7: 明示的話題の同定用辞書の例 (2)

まず第一に	次に
それから	ところで
あと	もう一つ

図 8: 明示的話題転換の検出用クルーワード辞書の例

3.2.3 階層 3: 関連話題

日本語における焦点候補は以下のような優先順位である: TOPIC、EMPATHY、主語、直接・間接目的語、その他 [W.Friedman *et al.*, 1990]。ここで、TOPICとは副助詞「は」でマークされた名詞句であり、EMPATHYとは「喜ぶ」「信じる」等の心理動詞の主語、「行く」の始点、「来る」の終点、受給表現「やる」「もらう」の主語、「くれる」の"に-格"などである。

関連話題としては、図9のような辞書を用いてTOPICやEMPATHY等を同定する。話題候補検出のためのパターン・マッチング処理と、話題に適さない不適切語に関する処理は明示的話題と同様である。

表層格	動詞
(話題) は	任意の動詞
(話題) が	喜ぶ、信じる やる、もらう
(話題) から	行く
(話題) に、(話題) へ	来る
(話題) に	くれる

図 9: 関連話題の同定用辞書の例

3.3 階層間のインタラクション

階層 1 において分野転換が検出されれば、階層 2 でも話題転換が起こったとみなされる。すると、たとえ階層 2 において話題転換が検出されていなくても話題転換が起きたとみなされ、例えば階層 2 における新しい談話セグメントが導入される。また、階層 2 において明示的話題転換が起きるかあるいは明示的話題が同定された場合には、階層 3 でも話題転換処理が行われる。

ここで問題となるのは、階層 2 と階層 3 における話題同定は即時に行われるのに対して、階層 1 における分野同定処理は過去 n 個の分野用語の頻度を用いて行なうので、実際に分野転換が起こってから検出されるまでには遅延が生じることである。このため、分野転換と明示的話題転換が同時に起きた場合には、階層 1 において分野転換が検出されて階層 2 においてそれに伴う話題転換処理を行おうとしても、その話題転換は既に階層 2 において処理済みであるという現象が生じる。

この重複を避けるためには、階層 1 で分野転換が検出された際に、分野転換が起きた本当の時刻を推定し、その時刻までさかのぼって階層 2 の状態を調べ、その上で階層 2 に分野転換情報を伝達するかどうかを決定する必要がある。実際に話題転換が起こった時刻を正確に推定するのは困難であるが、例えば分野同定用の FIFO の真ん中の分野用語が検出された時刻 t とすることができる。

3.4 処理例

図 1に示す対話を例とり、階層型話題同定モデルにおける処理例を説明する。階層 1 において分野同定に用いる FIFO の長さは 5 とする。対話が始まる前の初期状態において、FIFO は空であり、明示的話題と関連話題を管理するための各集合も空である。

発話 A-1 では分野用語として「郵便物」が検出される。これは図 5によると「郵便」分野の用語であるので FIFO 中の郵便分野用語数が 1 となる。この時点では FIFO 中の要素数が FIFO の長さ 5 より少ないので分野は同定できない。また、図 7の「(話題) について」という表現とマッチングがとれるので「郵便物の出し方」が明示的話題として同定される。発話 A-1 後の状態は図 10のようになる。

FIFO:	(郵便)
分野:	不明
明示的話題:	(郵便物の出し方)
関連話題:	()

図 10: 発話 A1 後の話題同定モデルの状態

発話 B1 でも分野用語として「郵便物」が検出される。発話 A2 では「国内便」と「ポスト」が順に分野用語として検出される。図 5によると対応する分野は、「国内便」が「郵便」と「航空」、「ポスト」が「郵便」と「ビジネス」である。また、図 9の「(話題) は」という表現とマッチングがとれるので、「国内便」が関連話題とし

て同定される。発話 A2 後の話題同定モデルの状態は図 11のようになる。

FIFO:	(郵便, 航空) (郵便) (郵便)
分野: 明示的話題: 関連話題:	不明 (郵便物の出し方) (国内便)

図 11: 発話 A2 後の話題同定モデルの状態

発話 B2 では処理は何も行われない。発話 A3 では「エアメイル」と「ポスト」が順に分野用語として検出される。最初の「エアメイル」が検出された時点でようやく FIFO 中の要素数が FIFO の長さ 5 に達するので分野の同定が可能となる。FIFO 中の分野用語数は「郵便」が最大であるので、現在の分野は「郵便」分野であると同定される。また、図 9の「(話題) も」という表現とマッチングがとれるので、「エアメイル」が関連話題として同定される。発話 A3 終了後の話題構造は図 12のようになる。

FIFO:	(郵便, ビジネス) (郵便) (郵便, ビジネス) (郵便, 航空) (郵便)
分野: 明示的話題: 関連話題:	郵便 (郵便物の出し方) (エアメイル, 国内便)

図 12: 発話 A3 後の話題同定モデルの状態

発話 B3 では処理は何も行われない。発話 A4 では「切手」が「郵便」分野の用語として検出される。また、「もう一つ」という表現が図 8に含まれることから、明示的話題転換が検出される。階層 2 における話題転換処理として、明示的話題を管理していた集合を空にリセットし、さらに話題転換に関する情報を階層 3 にも伝え、関連話題を管理していた集合も空にリセットする。また、7の「(話題) について」という表現とマッチングがとれるので、新しい談話セグメントにおける明示的話題として「切手の購入」が同定される。話題同定モデルの状態は 13のようになる。

FIFO:	(郵便) (郵便, ビジネス) (郵便) (郵便, ビジネス) (郵便, 航空)
分野: 明示的話題: 関連話題:	郵便 (切手の購入) ()

図 13: 発話 A4 後の話題同定モデルの状態

4 話題同定モデルの評価

4.1 実験方法

人手による話題同定の結果と比較することにより話題同定モデルの評価を行なった。対話データは文字起こししたものであり、「総務業務」と「プロ野球」に関するものの 2 種類を用いた。

1. 被験者による話題同定: 談話セグメントの区切りとそこでの話題の同定を 3 人の被験者に個別に行なってもらう。
2. 被験者へのインタビュー: 実験者が各被験者の話題同定結果をみて、確認や矛盾点に関する質問を行なう。この時点で、もし区切り等に矛盾がみつかれば訂正を行なう。
3. 話題同定の正解の作成: 3 人の被験者が同定した話題を元に正解を作る。正解版を学習用のものと評価用のものに無作為に振り分けた (図 14)。
4. 辞書の人手による学習 話題同定等に用いる辞書の項目の追加、修正、削除など。
5. 評価: 評価用データに関して、話題同定モデルによる同定結果と被験者による結果を比較する。

談話セグメントの区切り方や話題の同定に関する個人差はあまりないと予想されたが [J.Grosz & L.Sidner, 1986]、現実にはどうか分からなかったので最初は談話セグメントの区切りと話題同定の過程を分けた。幾つかのデータに対してセグメント区切りを行なってもらった結果あまり個人差がないことが分かったので、後半の対話データに対しては談話セグメントの区切りと話題同定を同時に行なってもらった。なお、個人差が目だったの

は、区切りの細かさと、話題が明確でない発話が続いた場合の区切り場所であった。

また、被験者へのインタビューは非常に有効であった。例えば、談話セグメントとしては2つに区切っていても、各セグメントにおける話題は同一のものとしている場合が何度かあった。このような場合インタビューを行なうと、同じ話題だということが意識されて1つの談話セグメントに統合されたり、話題の違いが認識された。

	学習用	評価用
総務対話	5対話、226文	4対話、314文
野球対話	1対話、45文	1対話、448文
合計	6対話、271文	6対話、1087文

図 14: 学習用と評価用の対話データ

4.2 人手による話題同定との比較

話題が正しく同定されている発話文の割合を各階層ごとに調べた。その結果を図15に示す。実際にシステムで処理した結果であるが、話題同定を行なう前の形態素解析の失敗等の要因は除き、純粋な話題同定モデルの性能を評価した。

分野用語がほとんど使われていない対話や、話題が明示的には導入されにくい対話もあった。このような対話における話題は現在の表層的処理では扱うことはできない。したがって、話題同定能力はある程度限られるが、話題に応じた情報提供等の応用は可能である。

	総務対話	野球対話
分野	295(94%)	448(100%)
明示的話題	193(61%)	87(15%)
関連話題	129(41%)	146(33%)
分野・明示的話題	193(61%)	87(15%)
分野・明示的話題 関連話題	47(15%)	10(2%)

図 15: 話題同定に成功した発話数

5 まとめ

談話を話題という側面から捉えた話題展開モデルを提案した。このモデルは分野、明示的話題、明示的話題転換、関連話題から構成される。従来の談話モデルでは目的志向型対話のみを対象としていたが、このモデルでは解説文のような目的志向でないような対話も扱うことができる。

次に、話題展開モデルの話題構造を逐次的に同定するための計算枠組の話題同定モデルを提案した。話題構造を階層的にとらえて同定を行う。

最後に、話題同定モデルによる結果と人手による話題同定結果を比較することにより話題同定モデルの評価を行なった。その結果、ある程度の精度が確認された。

本稿で提案した階層型話題同定モデルは「表層的言語表現」とのマッチングのような表層的処理のみを用いるので改善の余地は大いにある。今後は他の要素も考慮した話題同定モデルを考えていきたい。

参考文献

[J.Grosz & L.Sidner, 1986] Barbara J.Grosz and Candace L.Sidner. Attention, intention and the structure of discourse. *Computational Linguistics*, Vol. 12, No. 3, pp.175–204, 1986.

[J.Litman & Allen, 1987] D. J.Litman and J. F. Allen. A plan recognition model for subdialogues in conversations. *COGNITIVE SCIENCE*, Vol. 11, No. 1, pp.163–200, 1987.

[L.Sidner, 1983] Candace L.Sidner. Focusing in the comprehension of definite anaphora. In Michael Brady and Robert C.Berwick (Eds.), *Computational Models of Discourse*, pp. 267–330. MIT Press, Cambridge, 1983.

[W.Friedman *et al.*, 1990] Marilynn W.Friedman, Masayo Iida, and Sharon Cote. Centering in Japanese discourse. In *Papers presented to the 13th International Conference on Computational Linguistics*, Helsinki, Finland, August 1990.