自然言語処理 124－3
（1998．3．12）

主題間の関係を用いた文脈構造ネットワークの構築

吉田 悦子† 横山 晶一†† 西原 典孝††

†山形大学大学院 工学研究科
††山形大学 工学部

概要

文脈構造を正確に解析するためには主題間の関係を捉えることが重要である。本稿では、文中から主題・焦点を抽出した結果を用いて、文脈構造をネットワーク化する方法を提案する。その方法は、抽出された主題・焦点にそれぞれ分類番号を付与し、２文間の関係をネットワークで表現する。この２文間の関係には直列型、並列型、推移型、従属型、入れ子型、転換型という６種類のネットワーク関係を設定した。全ての２文間の関係を決定した後、談話全体を通して修正を加える。その結果、主題から話題の展開を捉えることができた。また、主題の抽出から文脈構造ネットワークの構築までの一連の動作を機械で一括して処理することが可能となった。

Construction of Context Structure Network Using Relation of Themes

Etsuko YOSHIDA†, Shoichi YOKOYAMA††, Noritaka NISHIHARA††

† Graduate Course of Engineering, Yamagata University
†† Faculty of Engineering, Yamagata University

Abstract

In order to analyze the context structure, it is important to catch the relation of themes. This paper proposes a method of constructing the context structure network using themes and focuses. The method firstly gives meaning class numbers to derived themes and focuses, and then represents a relation between two sentences as a network. Six network relation types are set up, that is, sequential, parallel, transitional, dependent, insert, and turning ones. After all relations are decided, the network relations are modified through the discourse. In the result, it is possible to process automatically deciding themes to constructing the context structure network.

1 はじめに

近年、大量の情報を高速で処理することが求められているが、これを行うためには、文章を機械処理して、話題の展開を自動的に把握すること、即ち文脈解析が必要である。これまでの文脈解析の研究は、名詞の出現頻度や照応関係、キーワード抽出の観点からが主体であった。しかしながら、文脈構造を正確に捉えるためには、主題を抽出して、そのつながりをみることが必要である。

そこで本研究では、文から主題を抽出することによって、文脈構造をネットワークで表現することを試みた。主題の抽出方法については既に報告している [吉田 97]。今回は、文脈構造ネットワークの構築方法を提案する [吉田 98]。

文脈構造ネットワークには、主題抽出方法で抽出した主題・焦点、接続詞、指示語というパラメータを用いた。まず、接続詞の情報から前後文の関係を決定し、次に抽出された主題・焦点とその修飾句の要素に分類番号を付与することにより、その番号から文間の文脈関係を明らかにする。その文脈関係をつなげることで、文章全体の構造をネットワークで表現した。その結果、文章から主題抽出を経て、文脈構造ネットワークの構築までの一連の動作を機械処理することが可能となった。

検証には朝日新聞 天声人語 [M-1]、毎日新聞データ集 [M-2] を用いた。このうち、本研究では天声人語については1993年を、毎日新聞については1994年のデータを用いた。これは内容の偏りをなくすため、通年にわたり極端な出来事がない年を選んだためである。

2 文脈構造ネットワーク

文章の構造には、書式的なものと内容的なものの２種類があげられる。書式的なものでは章、節などが、内容的なものでは起承転結などが考えられる [小野 89]。本研究においては、後者の内容的な部分の構造をネットワーク化する。

文脈構造とは文間の接続関係を記述したものである。この文脈構造を表現するネットワークが文脈構造ネットワークであり、本研究では次のように定義する。

文章全体を構成する各文の主題・焦点をもとに文章全体の構造を表現するネットワーク

また、文脈構造ネットワークを構築するには、次のパラメータを用いることが可能である [田村 97]。

1) 接続語
2) 指示語
3) 主題
4) 時制
5) モダリティ

これらは全て表層的な情報である。1) では「さて、ところで、では」といった接続語が話題転換の指標となる。2) では指示語が前文の名詞を指している場合などに、前文とその文に密接なつながりがあると考えられる。3) では主題は話題の中心となる要素なので、主題同士の意味的関係から接続関係を導くことができる。4) では前文とその文の時制が一致していない場合、話題転換の可能性が考えられる。5) ではモダリティ的表現により、話し手の状況などを見ることができる [仁田 91]。これにより、文の前後関係を推定することも可能である。

このうち本研究では文脈構造ネットワークを構築するために、1) の接続詞、2) の指示語、3) の主題、それぞれの情報を用いた。1)、2) は表層に現れている情報であり、文中から抽出することが容易である。また、主題よりも明確に文同士の接続関係を示しているので、より強い制約となる。

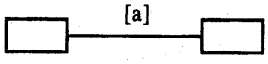
3 文脈構造ネットワークの構築

ここでは、文脈構造ネットワークを表現するためのネットワーク関係と、構築方法について説明する。

3.1 ネットワーク関係

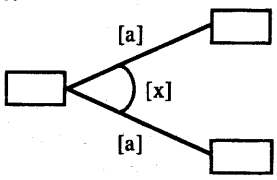
ネットワーク関係とは、２文以上の接続の仕方を表したものである。このネットワーク関係には、直列型、並列型、推移型、従属型、入れ子型、転換型の６種類がある。それぞれの型は次のような

1) 直列型

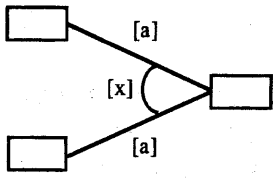


2) 並列型

a) 派生

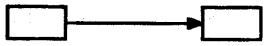


b) 収束

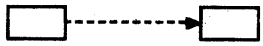


3) 推移型

a) 基本形

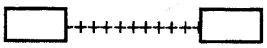


b) 異なる段落

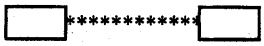


4) 従属型

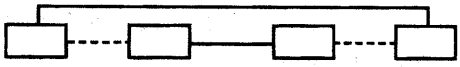
a)



b)



5) 入れ子型



6) 転換型

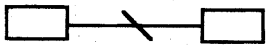


図1　ネットワーク関係図

ものである。

1) 直列型
話題が同一である２文間のネットワーク関係を表す。

2) 並列型
話題が１文から２文以上に派生するか、２文以上が１文に収束するという形式のネットワーク関係を表す。

a) 派生
１つの話題が２つ以上の話題に派生している形式

b) 収束
２つ以上の別の話題が１つの話題に収束する形式

3) 推移型
話題が全く異なるのではなく、何らかのつながりはあるが、話題が移行するというネットワーク関係を表す。

a) 基本形
前方の文の主題が後に続く文と接続していて、その文の焦点と後方の文の主題が意味的に一致しているもの

b) 異なる段落
段落で区切られているが後方の文の焦点が前方の主題・焦点と意味的に一致するもの

4) 従属型
同一段落中で後方文の焦点と前文が関係のある２文間のネットワーク関係を表す。

a) 同一段落中で前方の文の主題と後方の文の焦点のみが意味的に一致するもの
b) 同一段落中で前方の文の焦点と後方の文の焦点のみが意味的に一致するもの

5) 入れ子型
ある話題が別の話題を飛び越えて、再びその話題に戻るというネットワーク関係を表す。

6) 転換型
話題が全く別のもので、話題転換が起こっているというネットワーク関係である。
これは、前方と全くつながりが存在しない場合である。

以上の6つの型でネットワーク関係を表現する。それぞれの表現方法を図1に示す。

3.2 文脈構造ネットワークの構築

ここでは前節で述べたネットワーク関係を用いた文脈構造ネットワークの構築方法を示す。

3.2.1 文脈構造ネットワークの構築方法

文脈構造ネットワークの構築方法は以下の通りである。

1) 主題・焦点の抽出
入力されたそれぞれの文について主題・焦点を抽出する。

2) 分類番号の付与
抽出した主題・焦点とその修飾句内の名詞にそれぞれ分類番号を付ける。

3) 文脈構造ネットワークの決定
接続詞から前後文の接続関係を決定した後、後方文の主題との接続関係を決定し、段落情報を加えて修正し、ネットワークを構築する。

このような方法で文脈構造ネットワークの構築を行う。それぞれの方法について、次項で説明する。

3.2.2 分類番号の付与

文脈中では、主題や焦点を表す言葉が常に同じ語になるとは限らない。このような場合に主題・焦点の意味的なつながりをみるために、抽出した主題・焦点に分類番号を付与する。その分類番号は角川類語新辞典 [M-3] で検索し、上位3桁を採用する。

例3-1) 教室：分類番号941（カテゴリ：部屋）
絵：分類番号861（カテゴリ：絵画）

この3桁の分類番号が一致するとき、文の間につながりがあると解釈する。

また、主題と焦点には修飾句がかかっている場合が多く、修飾句の名詞も文脈に深く関係があるので、修飾句中の名詞にも分類番号を付与する。

例3-2) 満月の50倍も明るい第二の太陽
満月：分類番号005（カテゴリ：月）
50倍：分類番号262（カテゴリ：増減）
第二：分類番号183（カテゴリ：箇条）
太陽：分類番号004（カテゴリ：太陽）

複合名詞の場合も各々の名詞が文脈に深く関係することがあるので、各名詞に付けられた分類番号及び複合名詞全体に対する分類番号を全て採用する。

例3-3) 単身赴任
単身：分類番号128（カテゴリ：単複）
赴任：分類番号564（カテゴリ：従業）
単身赴任：分類番号564（カテゴリ：従業）

分類番号が複数ある場合、すべてを情報として与えておき、ネットワーク関係を決定するときに前後で接続するものを採用する。

一般主語「人」「人々」などを表す分類番号507が主題に存在した場合、修飾語があるならばその分類番号も加え、完全に一致しない限り、同一とみなさない。これは、例3-4) のように修飾語によって一般主語「人」の指す範囲が異なるためである。

例3-4) 見る人≠写す人

3.2.3 文脈構造ネットワークの決定

ここでは、文脈構造ネットワークの決定方法について説明する。

文脈構造ネットワークの決定方法

1) 接続詞処理（前後文の関係）

接続詞の情報からネットワーク関係を決定し、その情報に沿ったラベル付けを行う。

代表的な接続詞 [佐治 91] と対応するネットワーク関係、ラベルを示す。

a) 直列型
ラベル＝順接
そこで・だから・従って・それで
ラベル＝逆接
しかし・だが・ところが・けれども
ラベル＝説明
つまり・即ち・但し・たとえば

b) 並列型
ラベル＝並立・累加
また・そして・なお・さらに・そのうえ
ラベル＝対比・選択
または・あるいは・もしくは・それとも
但し、第何文と並列関係になるかは２文間の関係だけではわからないので、すべての関係が決定してからラベルを付ける。

c) 転換型（ラベルなし）
さて・次に・ところで・では・ときに

2) 主題によるネットワーク関係の決定

2-1) 前後文の場合
後方文の主題に代名詞が存在する場合は、次の代名詞処理を行う。
存在しないならば、2-2 の処理に移る。

・代名詞処理
第 i 文の主題を A、焦点を B、第 i+1 文の主題を代名詞 +C、焦点を D とする。

a) C = φ：直列型
b) C ≠ φ
b-1) A = C：直列型
b-2) A ≠ C：直列型
Aが他と接続していない
b-3) A ≠ C：推移型 a
Aが他と接続している

2-2) 第 i 文と第 j 文（前後していない文）の関係の決定
a) 主題 i = 主題 j：直列型
b) 焦点 i = 主題 j：
主題 i が他と接続していない：直列型
主題 i が他と接続している　：推移型 a
c) 上記以外：3) へ

3) 焦点によるネットワーク関係の決定

2) の処理が終了した時点で、どの文ともネットワーク関係が決定していない文が存在するとき、この処理を行う。

3-1) 段落内の文との関係を決定
a) 主題 i = 焦点 j：従属型 a
b) 焦点 i = 焦点 j：従属型 b
c) 一致せず：3-2) の処理を行う
3-2) 段落外の文との関係を決定
3-1で決定しなかった場合のみ、処理を行う
a) 主題 i = 焦点 j：推移型 b
b) 焦点 i = 焦点 j：推移型 b
c) 一致せず：主題・焦点の抽出誤り

4) 修正

談話全体を通して、修正する

4-1) 並列型の決定：
２文以上の接続がある場合
接続詞処理で保留されていた場合は、そのラベルを付け加える。
4-2) 入れ子型の決定：
入れ子構造になっている場合
4-3) 転換型の決定：
前方と全く接続していない場合
4-4) ラベル未定部分の決定：
直列型でラベルが未定のネットワーク関係は全て「同列」をラベルとする。

以上が、文脈構造ネットワークの構築方法である。

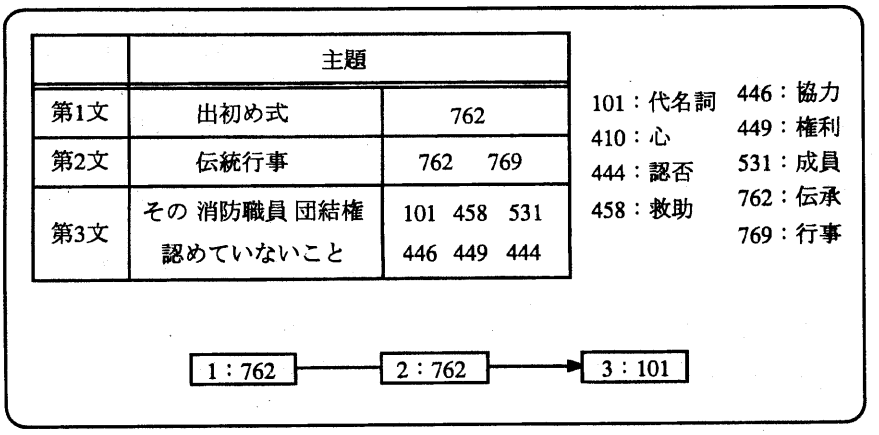


図2　直列型と代名詞処理のネットワーク関係図（文例3-6）

この方法に基づいて実際の文章で解析を行った結果を次節で示す。

3.3　解析

今回、天声人語20編、CD-毎日新聞'94データ集より10編について解析を行った。毎日新聞データは社説、社会面の記事、政治・経済の記事などを無作為に選んだ。それぞれのデータは6文から36文までのものである。これらの談話について主題・焦点の抽出を行い、その結果を用いた。ここでは、代表的な解析例を示す。例文中の下線部分は接続詞、イタリック＋下線の部分は主題、イタリック部分は焦点を表している。

例3-5）接続詞処理の例
段落5
　第14文
ことばで、実体と違うものを想像させることもできるし、理解しにくいように煙幕も張れる。
段落6
　第15文
例えば、医療の世界でドイツ語が長く使われてきた。

例3-5の談話では、接続詞「例えば」が存在する。このため、前文の主題が何であろうと、接続詞の情報からネットワーク関係を決定する。第14文と第15文のネットワーク関係は直列型、ラベルは「説明」である。

例3-6）直列型と代名詞処理の例
段落1
　第1文　*出初め式*は気持ちがいい。
　第2文　*伝統行事*は人の心を暖める。
段落2
　第3文　その消防職員に団結権を*認めていないの*は、先進国といわれる国では*日本だけ*である。

例3-6では、第1文と第2文の主題の分類番号が一致しているので、第1文と第2文は直列型である。第2文の主題「伝統行事」と第3文の主題「認めていないこと」は一致しないが、第3文の主題にかかっている修飾句に代名詞「その」があるため、代名詞処理を行う必要がある。この例では、第1文と第2文が直列型で第2文の主題が接続しているので、代名詞処理b-2に相当する。従って、第2文、第3文のネットワーク関係は推移型aとなる。この例の分類番号とネットワーク関係を図2に示す。

例3-7）並列型への修正の例
段落4
　第10文
1950年代後半に発表した*「パーキンソンの法則」*だ。
　第11文
簡単にいうと、どんな組織においても実際の仕事の量とは無関係に職員の数は毎年増えてゆく。

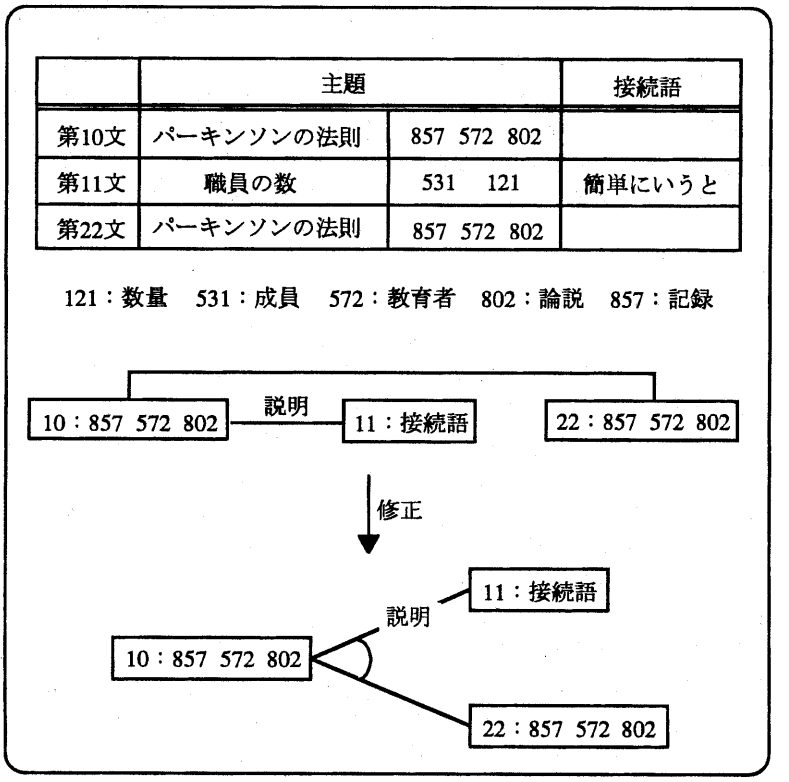


図3 並列型への修正（文例3-7）

段落7

第22文

「パーキンソンの法則」が出版された当時、特に中国で［φ（パーキンソンの法則）は］関心が高く、売れに売れて、上海では二日で売り切れたそうだ。

例3-7の談話では、第11文に説明の接続語「簡単にいうと」が存在している。このため、第10文と第11文は直列型になる。また、第10文と第22文の主題が一致しているため、第10文と第22文も直列型となる。従って、第10文から第11文と第22文が派生している形となり、並列型となる。このように、並列型は3文以上の関係から決定するため、談話内の全ての文に関してネットワーク関係を決定した後、並列型に修正しなければならない。

この例の分類番号とネットワーク関係を図3に示す。

4 結論

本研究では、談話解析の第一段階として、機械処理の観点から各文の主題の抽出方法をアルゴリズム化し、さらにその主題を用いた文脈構造ネットワークの構築を試みた。今回、天声人語20編、毎日新聞データ10編に対して文脈構造ネットワークの構築を行った。その結果、文脈構造ネットワークによって談話中の話題展開形式を表現することが可能となった。これにより、テキストデータから主題・焦点を抽出し、文脈構造ネットワークを構築するという一連の動作を機械処理することが可能となった。

現段階での主な問題点は次の通りである。

1）ネットワーク関係未定の問題

本研究では、焦点を用いても決定できない場合「主題・焦点の抽出が誤りである」とした。今回解

析を行った談話では、決定できなかった例について解析したところ、全て主題の抽出が誤りである場合であった。しかし、主題抽出は正解であるが、分類番号の分類が細かいために一致しない可能性もある。この場合、一段階上の分類を用いる必要があるのかどうかを考察する必要がある。

2) ラベル「説明」の問題

ラベル「説明」に属する接続詞の場合、直列型としたが、「説明」という性質上、前文の補足的な内容であることが多い。そのため、従属型ではないかと考えられる。しかし、これに相当する文が少ないため、さらに解析する必要がある。

3) 代名詞の取り扱い

代名詞が存在する場合は前文との関係のみを決定していたが、代名詞が前文を越える場合や、後方の文の要素を照応することがある。このため、前文との関係以外にも、主題が接続している文との関係についても考察する必要がある。

また、本研究では代名詞の指示対象までは言及しなかったが、主題の接続関係から指示対象を絞ることが可能ではないかと考えられる。これについても、今後解析する必要がある。

4) 分類番号一致の問題

今回の解析では、分類番号が1つでも一致する場合、ネットワーク関係を決定した。しかし、修飾句中の番号も全て一致する場合と、1つのみ一致の場合に接続の強さの違いがあるのではないかと考えられる。この違いを明確にし、違いがあるならば、どのように差をつけるのかを考察する必要がある。

これらの問題点を解決し、文脈構造ネットワーク構築までの一連の動作をシステム化することが今後の課題である。

参考文献

[小野 89] 小野顕司,浮田輝彦,天野真家：文脈構造の分析,情報処理学会研究報告,NL70-2,情報処理学会（1989）

[佐治 91] 佐治圭三：日本語の文法の研究,ひつじ書房（1991）

[清水 95] 清水一澄,横尾英俊：日本語理解システムのための視点抽出と照応解決,情報処理学会論文誌,Vol.36,No.2,pp.247-246（1995）

[田村 97] 田村直良,和田啓二：統合と分割による文章の構造解析,言語処理学会 第3回年次大会発表論文集（1997）

[仁田 91] 仁田義雄：日本語のモダリティと人称,ひつじ書房（1991）

[野田 96] 野田尚史：新日本語文法選書1「は」と「が」,くろしお出版（1996）

[吉田 97] 吉田悦子,横山晶一：主題・焦点を用いた文脈解析の一手法,電子情報通信学会技術報告,NLC97-29,電子情報通信学会（1997）

[吉田 98] 吉田悦子：主題の抽出と文脈構造ネットワークの構築,修士学位論文,山形大学大学院工学研究科（1998）

参考資料

[M-1] 朝日新聞論説委員会：朝日新聞 天声人語 '93 春・夏,原書房（1993）

[M-2] 毎日新聞社：CD - 毎日新聞 '94 データ集,日外アソシエーツ

[M-3] 大野晋,浜西正人：角川類語新辞典,角川書店（1994）