

説明的テキストにおける文の接続関係の解析手法

田中智博[†] 林良彦[†]

テキストを構成する文間の接続関係を把握しておくことは、テキスト全体の構造を解析したり、さらにその内容を理解する際の基礎となる。本論文では、操作マニュアル等に代表される説明的テキストを対象とした文の接続関係の解析手法を提案する。提案する手法は、テキストにより記述される世界の特徴、および、説明というテキストのタイプがテキストを構成する文のタイプ、および、その接続パターンを制約するという考えに基づいている。本手法では、まず文を構成する単位文とそれを結び付ける接続表現を基に各文のタイプを決定する。次に、実テキストから抽出した接続関係解析規則を適用して、隣接する2文間の関係解析を行う。接続関係解析規則は文タイプの並びによりインデックス付けされており、隣接する2文における各単位文間に関係付けを行う。ソフトウェアの操作マニュアルを対象とした評価実験によれば、本手法は隣接2文間の関係の約80%を正しく解析することができた。

Analysis of Cohesive Relations between Adjacent Sentences in Explanative Texts

TOMOHIRO TANAKA[†] and YOSHIHIKO HAYASHI[†]

It is important for structural analysis of texts and texts understanding to analyze cohesive relations between adjacent sentences. This paper proposes a new method to analyze cohesive relations between adjacent Japanese sentences based on constraints naturally contained in objective texts. We especially focused on explanative texts such as technical manuals. In these texts, the world to be explained constrains the sentence type and sentence adjacent patterns. For example, text for explaining a software command will describe some user operations and the associated system responses in some cohesive ways. We have developed a sentence type system suitable for software manuals and constructed a set of analysis rules. An experimental result shows that correct rate of our algorithm is about 80%.

1. はじめに

テキストあるいは談話を解析する手法として多くのものが提案されている。それらの手法を用いる手がかりに基づいて分類すると、大きく次の二つに分けられる。

- (1) 対象世界に関する先験的知識を特定の知識構造(スクリプト、意味ネットワークなど)を用いてあらかじめ記述しておき、それへの当てはめによりトップダウン的に解析を行うもの(例えば文献1)など。
- (2) 修辞表現や文のタイプといった言語表現から抽出される情報に基づいてボトムアップ的にテキストの構造を解析するもの(例えば文献2), 3)など。

前者の方法は、例えばテロ事件の報道記事などの特定のエピソード構造を持つテキストを対象としている。これは、対象世界に対する知識がテキストの深い理解を行うために重要であることを示すことに研究の主眼が置かれていたためである。一方後者の方法は、主に論説文章などを対象に、その構造化を目的として提案されてきた。これは、修辞的表現などの言語表現上の特徴から文のタイプを規定しやすいこと、さらに文のタイプを用いることにより論旨の展開に従った構造化が行いやすいことによると考えられる。

これらに対し我々の研究では、対象とするテキストのタイプをソフトウェアの操作マニュアル等に代表される説明的テキストに設定している。このような説明的テキストの量は、各種機器、ソフトウェアの普及に伴い確実に増加しており、その解析手法の確立は、計算機による新規文書の作成支援、既存のテキストからの知識情報の抽出などの応用の面からも重要である。

[†] NTT 情報通信網研究所
NTT Network Information Systems Laboratories

説明的テキストを対象とする場合、上記の(1)の手法には、説明的テキストにおいては明確なエピソード構造が存在しないため、先験的な知識を記述しておく、それを利用することが困難であるという問題がある。(2)の方法は、説明的テキストには明確な論旨展開が存在しないため、特定の修辞表現に基づく文タイプによる解析が困難な点が問題である。

また、テキスト構造あるいは談話構造の理論的枠組みに関する研究(例えば文献4)などは行われているが、そこでは、文書の作成支援、知識情報の抽出などの工学的応用に関しては述べられていない。

本論文では、以上のような問題意識の下で、説明的テキストの構造解析や理解の基礎となる文の接続関係の解析手法について提案する。2章においては、対象とするマニュアル文章の特徴について述べ、それに適した接続関係解析手法の考え方を説明する。3章では対象の特徴を利用した文タイプの決定について述べ、4章ではその並びに基づく接続関係解析手法の詳細について述べる。5章では、評価実験およびそれに対する考察を示す。

2. マニュアル文章を対象とした文の接続関係の解析

2.1 マニュアル文章の特徴

一般に多くの文章では、文章を構成する各文が密接な内容上のつながりを保ちつつ展開する⁵⁾。これらの内容上のつながりは、省略、照応、接続表現、語彙連鎖等の言語表現手段により達成される。これらの言語表現手段は、テキストの記述対象やテキストのタイプ等に影響されない普遍的なものであると考えられる。このような普遍的な手がかりに対し、本研究で対象とするソフトウェアの操作マニュアルにおいては、不特定多数の読み手が対象であり、書き手は読み手との間に想定される限られた共有知識に基づいて記述を行う必要がある⁶⁾ことから、記述される対象世界(この場合、あるソフトウェア製品)や、テキストのタイプ(この場合、操作説明)がテキストの特徴に反映されていると考えることができる。

そこで、典型的と考えられるソフトウェアの操作マ

* 特徴抽出マニュアル:

NEC MS-DOS ユーザマニュアル (117 段落: 545 文)
 JUST SYSTEM 一太郎ユーザマニュアル (77 段落:
 503 文)
 APPLE Macintosh ユーザマニュアル (70 段落: 490
 文)
 全部で約 250 段落, 約 1500 文。

ニュアルの文章*を対象とし、マニュアルテキストの特徴を分析した。その結果、以下のような特徴が得られた。

〈特徴1〉

「(ユーザが)～すると、(システムが)～する。」「(システムが)～するためには、(ユーザが)～する必要がある。」のように、主体としては、ユーザ(対象としているシステムを使用する人、マニュアルの読者)、システム(ユーザが操作する対象)である場合が多い。ただし、多くの場合、これらの主体は省略されている⁷⁾。

〈特徴2〉

「～したい時、～」(ユーザの希望を表す)、「～できる」(ユーザあるいはシステムの可能な動作を表す)、「～している」(ユーザあるいはシステムの動作状態を表す)、「～になる」(システムの変化状況を表す)等、使用されるモダリティ(助動詞、補助動詞等)によって、主体(ユーザ、システム)および、そのモダリティの伝達上の機能(希望、可能、動作状況、変化状況等)が明確に現れる場合が多い。

〈特徴3〉

テキスト全体の流れは、ユーザの操作手順、システムの動作順序に従って記述されており、「(システムが) ファイルを出力する。(システムが) ファイルを出力すれば、(システムが)～になる。」のように文の接続においても、前文で述べられた行為(動作)を引き継ぎながら、後文を述べていることが多い。

〈特徴4〉

図1の1に示すように、37%の接続文間において、単位文単位で何らかの関係をもっている。典型的な例を図2に示す。

〈特徴5〉

形式段落単位内で意味内容がまとまっており、形式段落の中は、以下に示すような修辞表現により、構成ブロックがマークされていることが多い。

a. 接続詞および相当表現

例:

「かな漢字変換する場合、FEP をオンにします。

.....

ただし、高速モードになっている時は、

かな漢字変換することができません。」

b. 特定の文タイプの組み合わせ

例:

「日付を変更する時、～して下さい。

.....

日付を変更しない時、～して下さい。」

c. 主題等の語彙の連鎖

例:

「日付は、半角で入力します。

半角モードにするには、～キーを押して下さい。

日付は、～コマンドで見ることができます。」

以上のような特徴から、ソフトウェアの操作マニュアルにおいては、対象とする世界（ソフトウェア製品）とテキストのタイプ（操作説明）が文章をゆるやかに制約していると考えられる。すなわち、対象世界の特徴から、文章を構成する文の主体は、ユーザ（読み手）またはソフトウェアシステム（説明対象物）であり（〈特徴1〉）、これらは、使用される動詞や叙述表現によりマークされる（〈特徴2〉）。さらに、テキストのタイプの特徴から、自然なテキストの流れが存在する（〈特徴3〉）。また、接続関係にある二文間においては、主に節単位で意味的な関係が存在する（〈特徴4〉）。これは、書き手が読み手との共有知識が最小限しかない場合を想定し、できるだけ緻密に文章展開を進めようとするものの現れと考えられる。

2.2 マニュアルテキストの構造解析と接続関係解析

前節に述べたように、操作説明のマニュアルテキストには固有の特徴が存在する。すなわち、対象世界とテキストのタイプがテキストを構成する文のタイプとその接続の仕方を制約していると考えられる。したがって、マニュアルテキストの構造解析においては、このような特徴を活かした文の接続関係の解析が中心的役割を果たしうると考えられる。

テキストにおける文の接続関係解析については、桃内がサーベイを行っており⁸⁾、接続関係解析に関する基本的な問題として以下に示す3点をあげている。以下では、この3点に従って、我々の接続関係解析の考え方を説明する。

(1) 接続関係として何を設定するか: 桃内は、接続関係を設定する単位と設定される接続関係について

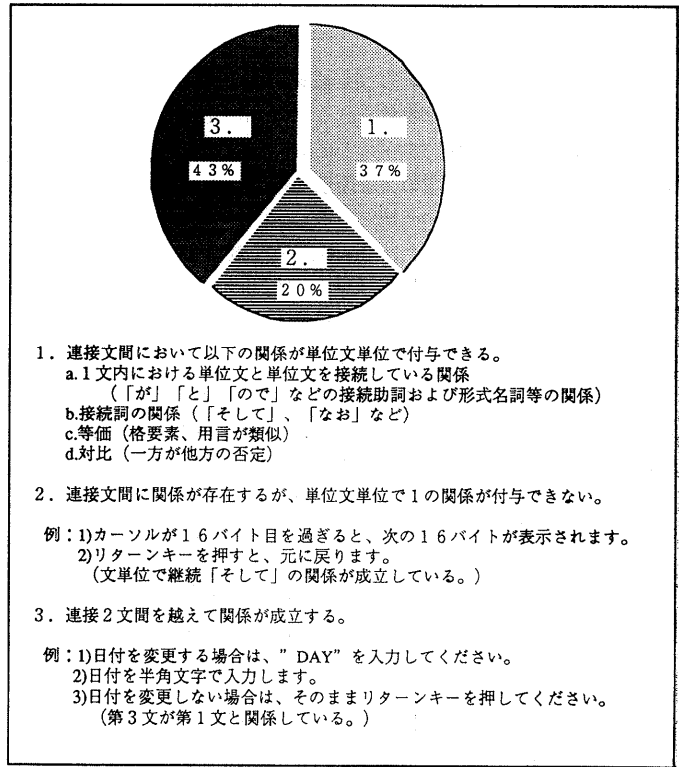


図1 マニュアル文章の接続文間の関係
Fig. 1 Relations between adjacent sentences in technical manuals.

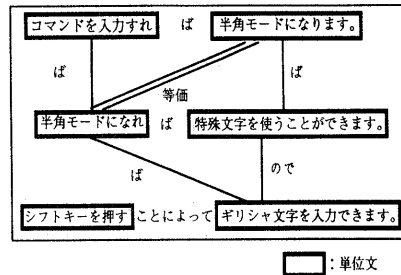


図2 接続文間の関係例
Fig. 2 An example of relations between adjacent sentences.

考察している。前者については、節、文、文群が基本的な単位であるとしている。また、後者については、接続関係、伝達的態度、修辭的關係、イベント間の時間・因果関係などをあげている。論説文章の構造化などにおいては、文が基本的な単位であり、設定される接続関係は修辭的關係となっていることが多い²⁾が、マニュアルテキストにおいては、前節に示した〈特徴4〉のように、ほぼ節に対応する単位文を単位とする意味的な接続関係が存在する。そこで本研究では、単

位文*を基本単位とし、それらの間に次章で述べる接続属性（1文内において単位文と単位文を接続している接続助詞等を分類したもの）、接続詞の関係（「そして」、「なお」等）、等価（格要素、用言が類似）対比（一方が他方の否定）の接続関係を設定する。

(2) 接続関係の解析をどのように行うか：桃内は、接続関係解析の手がかりとして、文の種類（平叙文、疑問文など）、文の意味内容の型（命題、事象など）、文要素間の関係などをあげている。本研究においては、前節の〈特徴3〉に注目した解析を行う。すなわち、〈特徴1〉、〈特徴2〉を反映した文のタイプを設定し、〈特徴3〉に従って、文のタイプの並びを基にした接続解析を行う。

(3) 接続関係の解析を文章全体の構造解析とどう統合するか：前節の〈特徴5〉に示したように、マニュアルテキストにおいては一つの形式段落は一つの意味的なまとまり（桃内の用語で言えば文群）をなす場合がほとんどであり、まとまりの中での話題の対比や転換は、接続的表現などの修辞表現によってマークされることが多い。そこで、接続関係の解析と修辞表現を解析することによる大局的な構造化を用いる我々のテキスト解析の構成を図3に示す。

このように、文の接続関係とより広い範囲を考慮し

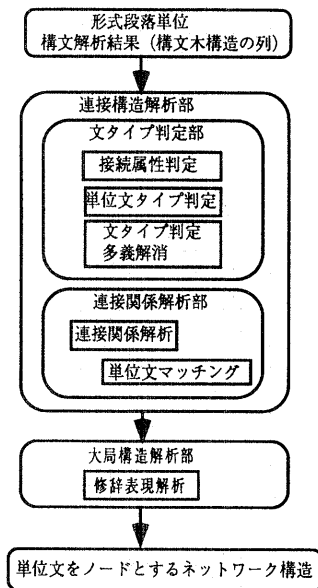


図3 処理の概要
Fig. 3 Analysis flow.

* 単位文：= 一つの述語表現と、その格要素、副詞要素からなる構文構造上の単位。

た修辞表現に基づく大局的な構造化を考慮することにより、形式段落内のテキストの構造の大まかな部分を把握できる。6章で簡単に述べるように、抽出された構造は、テキストからの知識情報のような応用の基礎となりうると考えられる。

3. 文タイプの決定

3.1 基本単位

対象世界の特徴と文章タイプの特徴に基づき文のタイプを分類する。ここでの分類は、解析の単位である単位文のタイプと、1文内において単位文と単位文を接続している接続助詞等を分類して得られた接続属性との組み合わせにより行う。1文内で二つの単位文が接続助詞等の接続表現によって接続されている場合（全体の約60%を占める）を基本形として、文のタイプを以下のように定義する。

文タイプ ==

(接続属性 従文単位文タイプ 主文単位文タイプ)*

また、1文内に単位文が単体で存在する場合には、接続属性を「単文」とし、次のように表す。

(「単文」 単位文タイプ)

そして、1文内に三つ以上の単位文が存在する場合には、上位の係り受け関係にある二つの単位文で代表させて、上記の形式にする**。

表1 接続属性分類

Table 1 Classification of connection categories.

接続属性	主な接続助詞, 接続表現
AFTER	「てから」
ANOTHER	「ほか」
BECAUSE	「ので」、「から」
BUT	「が」、「けれども」
BY	「て」、「により」(形式名詞に付く場合)
CONTINUE	「ながら」、「つつ」
IF	「ば」、「たら」、「なら」
OR	「か」(用言につく)
TIME	「たびに」
TO	「のに」、「ために」、「には」
UNTIL	「まで」、「かぎり」(用言につく)
WHEN	「と」、「時」、「場合」

* ここでの従文、主文は1文内における単位文の位置関係（接続属性をはさんで、前方にある単位文を従文単位文、後方にある単位文を主文単位文とする。）を表したものであって、意味上、主文が主であり、従文が従であるという関係を示したものではない。

** 残りの単位文は、係り受け関係に基づいて、代表としてあげられた単位文との間に関係（接続属性）を付与しておき、接続関係解析には使用しない。

3.2 接続属性の決定

接続属性の分類体系を表1に示す。この分類体系に従って、接続助詞等の表現から接続属性を決定する。ここで用いている分類体系は、「～すれば、～する。」(条件), 「～することによって、～する。」(方法)等、マニュアル文でよく用いられている用法を考慮して、通常よく用いられている体系^{9),10)}に若干の修正を加え、表面的な言語表現に近いレベルで分類したものである。

3.3 単位文タイプの決定

文の主体、機能を主な観点として、単位文のタイプを決定する。マニュアル文の特徴で述べたように、主体は、ユーザあるいは、システムである場合が多く、モダリティの部分に伝達上の機能が現れる。

主体として何を取りうるかという観点から動詞の分類を行い、その結果とモダリティ表現の解析を基にして、弁別ネットワークにより単位文のタイプを決定す

る。現在用いている動詞の分類と弁別ネットワークおよび、単位文タイプ分類体系を、表2、図4、表3に示す。

ここで、単位文タイプの主体の決定は、主に動詞の分類に依存している。しかし、動詞の分類だけではどうしても単位文タイプを一意に決められない場合がある(例えば、「～を行う」の主体がユーザであるのか、システムであるのか)。次にその解決策を述べる。

表2 動詞タイプの分類
Table 2 Classification of verb types.

動詞タイプ	対応する主な表現
「ユーザ」(ユーザの操作を表す)	押す, 入力する
「システム」(システムの動作を表す)	出力する, 表示する
「状態」(状態を表す)	消える, 現れる
「変化」(変化を表す)	変わる, 変化する
「その他」(一般的な動作)	行う, 使用する

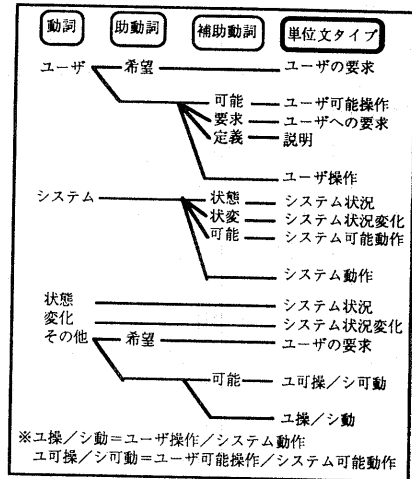


図4 単位文判定弁別ネットワーク
Fig. 4 Discrimination network for decision of unit sentence type.

表3 単位文タイプ分類
Table 3 Classification of unit sentence types.

単位文タイプ	定義 (例文)
ユーザ操作	ユーザ自身が実際に行う操作を述べた文 (コマンドを入力します)
ユーザ可能操作	ユーザが行うことのできる操作を述べた文 (スタックを開くことができます)
ユーザへの要求	システム、筆者のユーザへの要求を述べた文 (ウインドウを閉じて下さい)
ユーザの要求	ユーザのシステムに対する要求を述べた文 (画面の表示を変えたい)
システム動作	システム自身が実際に行う動作を述べた文 (ファイル名を表示します)
システム可能動作	システムが行うことのできる動作を述べた文 (短時間にデータを読み取ることができます)
システム状況	システムが現在置かれている状況およびシステム全体の状態を述べた文 (カードには ID 番号が付いています)
システム状況変化	システムの状況の変化を述べた文 (カードの並び方が変わります)
説明	コマンド、装置、その他システムについて解説している文 (ファイル名のパラメータはコピー先の指定です)

3.4 文タイプの多義解消

文タイプを構成する三つの要素（接続属性、従文単位文タイプ、主文単位文タイプ）のうち、従文単位文、主文単位文のどちらか一方で多義が生じた場合、残りの要素（接続属性と一意に判定された単位文のタイプ）から、ソフトウェアマニュアルに特有の表現（例えば、「ユーザがある操作をする」と、システムがある動作をする。」などの表現）、接続助詞の機能に基づいた規則¹⁰⁾（例えば、「ながら」「つつ」で接続された二つの単位文は同主体である。）を用いて、多義の絞り込みを行う。

図5に示すように、「検索を行ないます」で、多義（ユーザ操作か、システム動作）が生じた場合、残りの要素（「と」が接続属性 WHEN であり、「検索コマンドを入力する」がユーザ操作である）から、「ユーザがある操作をすると、システムがある動作をする。」というソフトウェアマニュアルに特有の表現を使って、「検索を行ないます」の動作主をシステムに絞り込み、単位文タイプをシステム動作とする。

なお、この方法では、本論文における文タイプ判定手法において、従文単位文、主文単位文の両方で多義が生じた場合、多義を絞り込むことが困難である。この場合に対処するためには、接続属性だけの情報から多義を絞り込む方法、接続する文の情報により絞り込む方法等を検討する必要がある^{*}。

4. 文タイプの並びに基づく接続関係解析

4.1 解析手法

上記の文タイプを基に、出現するすべての隣接する文タイプの並び（以後、接続パターンと呼ぶ）に対して、その文内の各単位文間の関係（接続関係）を、解析規則として、あらかじめ記述しておけば、それを用いて文タイプ判定部で得られた接続パターンから接続文間の関係を求めることができる。本論文における接続関係解析の一例を図6に示す。ここで問題となるのは、果たして接続パターンだけで接続文間の関係を一意に決定することができるのか、あるいは、出現するすべての接続パターンを記述することができるのか、ということである。ここでは、これらの問題に対し

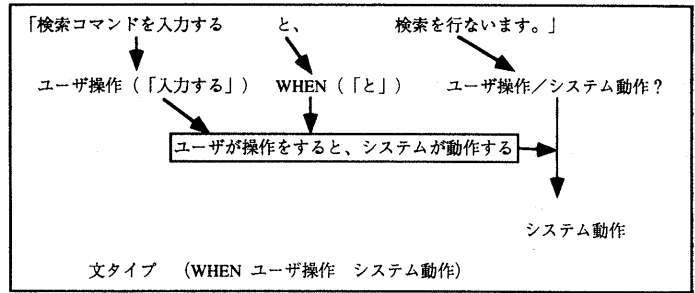


図5 文タイプ多義の絞り込み
Fig. 5 Disambiguation of unit sentence type.

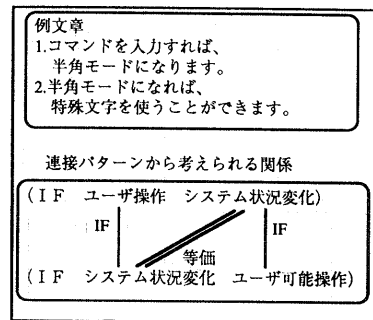


図6 接続パターン具体例
Fig. 6 An example of adjacent patterns.

て、次項で前者に対する解決策として、関係の多義の絞り込みの一手法を提案する。また、後者の問題に対しては、5章の考察で、接続パターンを基に記述した解析規則の収束性について議論する。

4.2 単位文マッチング

接続パターンだけからでは、図7に示すように同じ接続パターンでも接続関係が異なる場合が存在する。しかしながら、図7の例に見られるように、前文（文の並びにおいて先に現れた文）と、後文（文の並びにおいて後に現れた文）の間の接続関係のうち、一つの関係をダイナミックに決定することにより、接続文間の関係を一意に絞り込むことができる場合がある。図7の例では、前文の従文単位文（ユーザ操作）と、後文の従文単位文（ユーザ操作）の間の等価性と、前文の主文単位文（ユーザ可能操作）と、後文の従文単位文（ユーザ操作）の間の等価性のどちらかが判定できれば、接続文間の関係を図7の(A)の接続関係か、(B)の接続関係に決めることが可能となる。

ここでは、このように特定の単位文間の関係をダイナミックに決定する手法を、単位文マッチングと呼ぶ。単位文マッチングのアルゴリズムを図8に示す。入力は、その間の関係を判定したい2単位文とし、用

*ただし、調査、実験を行ったマニュアルテキスト中には存在しないため、現状では対処していない。

言および、格要素のマッチング部において類義、対義のデータによりマッチングを行う。それぞれの結果を結果判定部において統合し、等価、対比の関係を出力する。また、用言マッチング部において、マッチングが失敗した場合には、用言と格要素の組み合わせによ

るマッチング（「バックアップをとる」＝「コピーをする」等）を行い、マッチングによって得られた関係（等価、対比）を出力する。現在のところマッチングの判定の多くは、用言、格要素（名詞）、およびそれらの組み合わせに対するあらかじめ用意した類義、対義のデータに依存している。現状では、これらのデータは、実際のマニュアル文章から人手により抽出している*。今後は、意味属性などの辞書情報の有効利用や、データの半自動生成などについても検討する必要がある。

4.3 解析規則

解析規則のフォーマットは、図9に示すように、連接パターンをインデックスとし(α)、関係多義が存在する場合には単位文マッチングの結果を条件として、単位文間に接続関係を記述する(β)。なお、この部分は、LISPのcond節と同様の構文である。これにより、単位文間を結ぶリンクが張られ、そのラベルとして接続関係名が付与される。また、その規則内のすべての単位文マッチングが失敗した場合は、単位文間にリンク(ラベル: DEFAULT)を付与する(γ)。

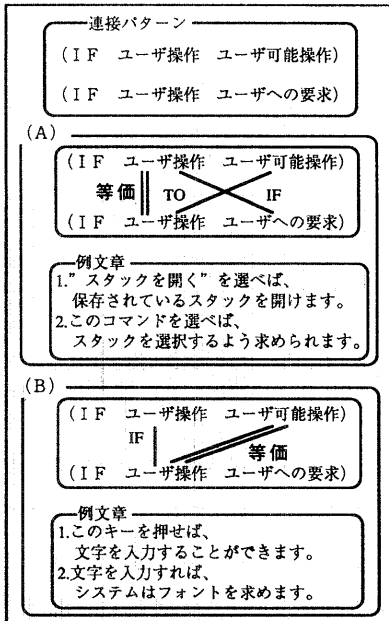


図7 連接パターン具体例
Fig. 7 An example of adjacent patterns.

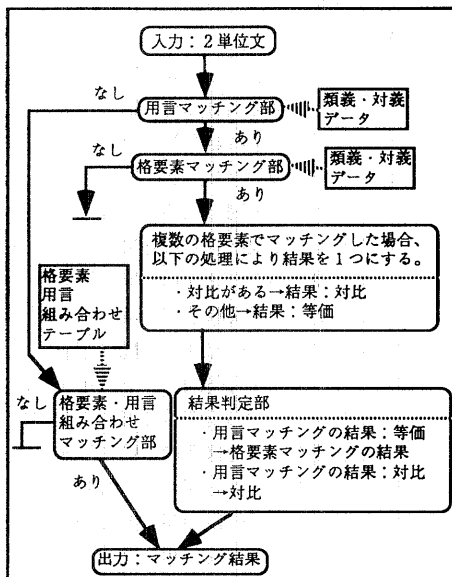


図8 単位文マッチングアルゴリズム
Fig. 8 Algorithm for unit sentence matching.

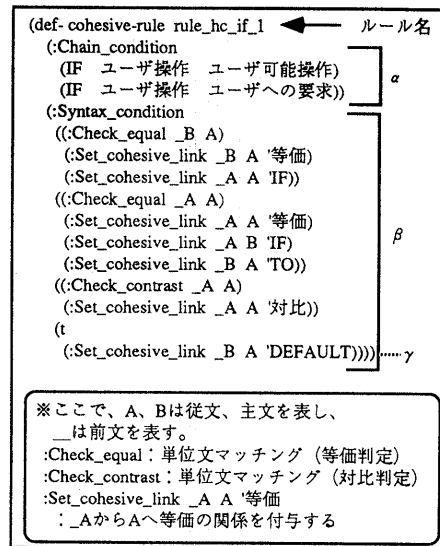


図9 解析規則一例
Fig. 9 An example of analysis rule.

* マニュアル文章約 1500 文から抽出した各データの数は、用言の類義語データ「選ぶ・選択する」など 48 語、対義語データ「指定する・省略する」など 6 語、名詞の類義語データ「ディスプレイ・画面」など 6 語、対義語データ「前・後」など 10 語、用言と格要素の組み合わせデータ「バックアップを取る・コピーを作る」など 6 対となっている。

5. 評価・考察

5.1 評価実験

評価は、本手法の精度および、解析規則の収束性を評価するために、次に示す二つの方法で行った。

一つは、文タイプの判定精度、解析規則による関係付けの精度のみを、他の条件（解析規則がすべての接続パターンをカバーしているか等）に影響されることなく評価するため、あらかじめ解析を行うマニュアル文章のすべての接続パターンを抽出し、その接続パターンを基に作成した解析規則を用いて実験を行う（クローズドデータ）ものである。

もう一つは、解析規則の収束性等を評価するために、解析規則を作成するためのマニュアル文章と、実験を行うマニュアル文章を分けて実験を行う（オープンデータ）ものである。次に各実験における結果を述べる。

5.1.1 クローズドデータによる評価実験

3種類のメーカーの異なるソフトウェアマニュアルより無作為に抽出した60段落からすべての接続パターンを抽出し、その接続パターンを基に解析規則を作成した。作成した規則数は、236で、このうち接続パターンによる制約のみから接続関係が一意に決まる（単位文マッチングを必要としない）規則数は、141であった。

また、あらかじめ机上検討により実際のマニュアル文章に対して人手により付与した文タイプおよび、接続関係を正解データとした。次に評価指標と結果を示す。

(1) 文タイプの判定

人手により判定した文タイプと一致した割合（文タイプ判定正解率）と、文タイプの多義を正しく解消できた割合（文タイプ多義解消率）を求め、それぞれ約80%、約87%が得られた。

$$\text{文タイプ判定正解率} = \frac{\text{人手により判定した文タイプと一致した文数}}{\text{実験に用いた文の総数}}$$

$$\text{文タイプ多義解消率} = \frac{\text{多義が正しく解消できた文数}}{\text{多義が生じた文の総数}}$$

(2) 接続文間解析

正しく文タイプが判定された文に対し、接続関係がどの程度正しく判定されたかをみるために、人手で判定した接続関係と、解析によって付与された接続関係

の一致した割合（規則適合率）を求め、約81%が得られた。また、単位文マッチングの効果をみるために、単位文マッチングを行わずに一番目の候補を無条件に付与した場合の規則適合率は約55%であった。これにより、単位文マッチングの効果を確認することができた。

規則適合率＝

$$\frac{\text{人手で付与した接続関係と一致した接続パターン数}}{\text{解析規則が適用された接続パターン数}^*}$$

5.1.2 オープンデータによる評価実験

2種類のメーカーの異なるソフトウェアマニュアルより無作為に抽出した264段落（1548文）を基に、規則作成用文数を変えて、出現頻度2以上の接続パターンについて解析規則を作成した。作成された規則数は、68（471文）、136（1104文）、166（1548文）である。この規則を使って、規則を作成したものとは異なるソフトウェアマニュアルより無作為に抽出した75段落（379文）を解析した。クローズドデータ同様、あらかじめ机上検討により、解析を行った75段落のマニュアル文章に対して人手により付与した接続関係を正解データとした。次に評価指標と結果を示す。

(1) 接続文間解析

クローズドデータで用いた規則適合率の結果は、約86%（作成用文数471文）、約86%（作成用文数1104文）、約88%（作成用文数1548文）であった。クローズドデータの結果と同様、解析対象文章を変更しても80%以上の解析適合率が得られることを確認した。

次に、作成した解析規則の適用により、対象文章がどの程度正しく解析されたかをみるために、人手で付与した接続関係と一致した接続パターンと、人手で接続関係を付与した接続パターンの総数との割合（解析正解率）および、作成した規則の適用精度を調べることを目的とする。解析規則が適用された接続パターン数と、人手で接続関係を付与した接続パターンの総数との割合（規則適用率）を求めた。結果を図10に示す。

解析正解率＝

$$\frac{\text{人手で付与した接続関係と一致した接続パターン数}}{\text{人手で接続関係を付与した接続パターンの総数}}$$

規則適用率＝

* クローズドデータの場合、机上検討により関係付けた接続パターン数と、解析規則が適用された接続パターン数は、同数である。

解析規則が適用された接続パターン数
 人手で接続関係を付与した接続パターンの総数

5.2 考 察

文タイプ判定, 接続文間解析, 解析正解率の収束性の各実験結果に対する考察を述べる.

5.2.1 文タイプ判定について

単語切り, 品詞の判定等形態素レベルでの失敗を除けば, 「れる, られる」(受身, 可能, 自発, 尊敬)等の意味に多義のある語の判定ミスである. これらの多義は文解析の段階で精度よく解消しておく必要がある.

また, 文タイプの多義の解消に失敗したものは, すべて単文(1文内に単位文が単体で存在している)である. 本論文における単位文タイプの多義の解消方法は, 1文内における接続属性および, 他の単位文タイプに依存しているため, 単文でタイプに多義が生じた場合に対処することは難しい. 接続属性や, 単位文タイプ以外の他の情報から判定する方法を検討する必要がある.

5.2.2 接続文間解析について

クローズドデータ, オープンデータとも接続パターンに単文が含まれる場合, 単文の文タイプ中の情報(接続属性等)による制約が欠落しているために, 解析規則作成時にその接続パターンに対する正しい関係のすべてを記述しきれていない.

例 接続パターン:

(単文 システム状況)

(WHEN ユーザ操作 ユーザへの要求)

- (1) ファイル名は半角入力になっています.
作成したファイルをセーブするときは,
半角モードにしてください.
- (2) ツールはディスクの中に保管されています.

ツールを使用するときは,
作業領域を確保してください.

例に示す接続パターンは, (1)に示すような「システムがある状況にあるので, ユーザがある操作をしたときは, ユーザが何かをしなければならない」というように前文が後文の理由を表している場合と, (2)に示すように, 前文で「あるシステムがある」ということを述べ, 後文では「そのシステムを使う」ことを述べている場合がありうる. このような場合に対処するには, 単文に対する文タイプ分類を細かく設定し, それに応じた解析規則を作成する必要がある.

また, 現状の単位文マッチングでは, 等価, 対比以外の関係を判定できないため, 上記の例のように, 接続関係中に等価, 対比の関係がない場合には適用できず, 多義を絞り込めていない. 単位文マッチングに対する考察を次に述べる.

5.2.3 単位文マッチングについて

例1: 「バックアップをとる」

= 「コピーをする」

例2: 「このコマンドを実行する」

= 「ボタンを削除する」

例3: 「ユーザの要求が高まる」

? 「ユーザの要求に答える」

例1のように, 文脈によらず等価と判定できるものに対しては, 類義関係として定義することにより, 表層上から判定が可能であり, 本論文で提案した単位文マッチングの手法が有効である. しかし, 例2のように表現の等価性が文脈に依存するものや, 例3のように現在のマッチングでは関係を判定できないものに対しては, 結果として出力する関係を含めた単位文マッチング手法の拡張が必要である. その際, 実データから抽出した知識の利用¹¹⁾, 省略や照応の解析, 対象世界における背景知識を援用した推論¹²⁾についても検討していく必要がある.

5.2.4 解析正解率の収束性について

図10に示すように, オープンデータによる実験の結果, 規則作用文を増やすことにより, 解析正解率が向上し, 規則作用文 1548 文で, 約 56% が得られた. 規則適合率が規則作用文の文数によらずほぼ一定の値(約 80%)であるため, 解析正解率の変化は, 規則適用率に依存していることがわかる. 今後, 解析正解率を向上するためには, 先に述べた解析規則の整備および, 単位文マッチングの高性能化による規則適合率の向上と, 規則適用率の向上が必要である.

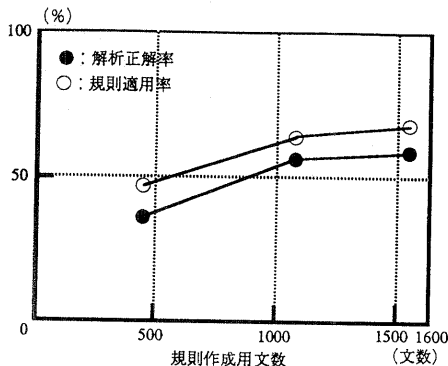


図10 オープンデータによる評価結果

Fig. 10 Results of the evaluation (open data).

次に規則適用率の向上について述べる。

オープンデータによる解析を行ったマニュアル文章 (379 文) 中, 人手により関係付けた 108 接続パターンおよび, 規則作成用文章 (1547 文) から抽出した 1284 接続パターンの出現頻度別の分布状況と, 各規則作成用文章 (471 文, 1104 文, 1548 文) における接続パターン出現頻度別の解析規則の規則適用率を図 11 に示す。図 11 に示すように, 出現頻度の高いものについては, 少数の規則作成用文章から, 規則を抽出することが可能である。すなわち, マニュアルの世界で共通して高頻度で使用される接続パターンが存在する。しかしながら, 全体の規則適用率を向上するためには, 低頻度 (現状では頻度 1) で現れる接続パターンに対して, 個々の解析規則を作成する必要がある。それらを規則作成用文から抽出するためには, 多くの文を必要とするばかりでなく, それらをすべて人手により処理することは困難である。

解析規則の整備, 単位文マッチングの高性能化による規則適合率の向上により, 解析正解率は規則適用率に接近し, 図に見られるように, 約 70% 程度まで向上可能である。さらに解析正解率を上げるためには, 低頻度で現れる接続パターンに対して, 解析規則の抽象化および効率的な作成方法の検討により, 解析規則を自動的に作成し, 自動的に関係を付与する機構を構築することによって, 規則適用率の向上を図る必要がある。

6. おわりに

本論文では, 操作マニュアルに代表される説明的テキストを対象とし, 記述される世界の特徴とテキストのタイプに着目した文の接続関係手法を提案した。提案する手法は, このような特徴がテキストを構成する文のタイプおよび文の並びをゆるやかに制約するという仮定に基づいている。クローズドデータに対する評価実験から, 提案する手法の基本的な有効性を示すことができたが, さらに精度を高めるためには指示詞の文脈機能の解析や背景知識を援用した推論が必要であることがわかった。ただし, これらの処理は高い処理コストを持つため, 必要に応じて選択的かつ部分的に

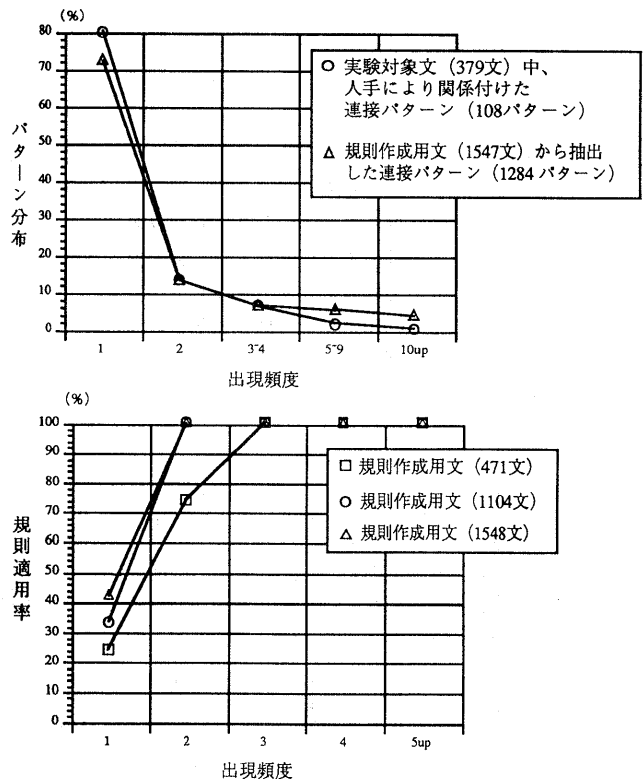


図 11 接続パターンの出現頻度と分布状況, および規則適用率の関係

Fig. 11 Relations between adjacent patterns frequency and applicative rate of the rules.

処理を起動する機構の開発が必要であろう。一方, オープンデータに対する評価実験からは, 解析規則のカバー率を向上させるために, 解析規則の抽象化や半自動的な規則獲得についても研究していくべきであると考えられる。また, 修辭的表現の解析に基づく大局的な構造化との統合によるテキスト構造解析の全体を実現し, 評価を進めていく必要がある。

以上に述べたように, 本論文で提案した文の接続関係解析は多くの課題を残してはいるが, テキスト構造解析の主要な一部をなすものであり, テキストからの知識情報の抽出などの応用の基礎となるものである。以下, この点について簡単に触れておく。各種機器やソフトウェアの普及によりマニュアルテキストの量は確実に増えていることから, 知的な検索システムのニーズが高まるものと考えられる。例えば, ユーザがある機能を実行したいがそのための操作手順を知らない場合に, 機能からそれらの実行に必要な操作手順を検索するシステムが考えられる。このためには, 機能

実行の必要条件となる操作と十分条件となる操作を判別し、マニュアルテキストの構造化を行っておく必要がある。本論文で提案する接続関係解析はそのための基礎となりうるが、現在設定している接続関係が表層的な言語表現に近いレベルであるため、目的・手段、あるいは、事象の因果関係を考慮したより意味的な関係へと変換する機能が必要になる。今後は、最近の指示表現の理解に関する研究^{13),14)}などを参考にしながら、上記の応用への展開も図っていく予定である。

謝辞 日頃よりご指導いただく中村稔メッセージシステム研究部部长、坂間保雄グループリーダーならびに同研究部の皆様方に感謝いたします。

参考文献

- 1) 石崎, 井佐原, 徳永, 田中: 文脈と対象世界モデルを利用した機械翻訳へ向けて, 人工知能学会誌, Vol. 4, No. 4, pp. 660-669 (1989).
- 2) 木下, 小野, 浮田, 天野: 日本語テキスト理解における文脈構造抽出法, 情報処理学会談話理解モデルとその応用シンポジウム, pp. 125-136 (1989).
- 3) 福本, 安原: 文の接続関係に基づく文章構造解析, 情報処理学会自然言語処理研究会報告, 88-2 (1992).
- 4) Grosz, B.J. and Sidner, C.L.: Attention, Intentions, and the Structure of Discourse, *Computational Linguistics*, Vol. 12, No. 3, pp. 175-204 (1986).
- 5) 市川: 国語教育のための文章論概説, 教育出版 (1978).
- 6) McGehee, B.M. (テクニカルライティング研究会訳): ユーザマニュアル執筆ガイド, p. 226, 日経マグローヒル (1987).
- 7) 林, 千葉: 日本語受動文の能動化可否判定アルゴリズムの検討, 情報処理学会論文誌, Vol. 31, No. 10, pp. 1438-1443 (1990).
- 8) 桃内: 文章における接続関係の解析のための基礎的考察, 情報処理学会自然言語処理研究会報告, 78-13 (1990).
- 9) 寺村: 日本語の文法(下), p. 153, 国立国語研究所 (1981).
- 10) 南: 現代日本語の構造, p. 325, 大修館書店 (1974).
- 11) 工藤, 樽松: 対話翻訳システムのための文脈処理機構とその性能評価, 情報処理学会論文誌, Vol. 33, No. 2, pp. 141-152 (1992).
- 12) Hobbs, J.: Coherence and Coreference, *Cognitive Science*, Vol. 3, No. 1, pp. 67-90 (1979).
- 13) Di Eugenio, B.: Understanding of Natural Language Instructions, The Case of Purpose Clauses, *Proc. of the 30th Annual Meeting of the ACL*, pp. 120-127 (1992).
- 14) Balkanski, C.T.: Action Relations in Rationale Clauses and Mean Clauses, *Proc. of COLING-92*, pp. 267-273 (1992).

(平成5年8月19日受付)

(平成6年1月13日採録)



田中 智博 (正会員)

1964年生。1987年大阪大学工学部産業機械工学科卒業。1989年同大学院工学研究科産業機械工学専攻修士課程修了。同年日本電信電話(株)入社。現在 NTT 情報通信網研究所メッセージシステム研究部研究主任。日本文推敲支援システム、情報検索システムなどの自然言語処理の研究に従事。人工知能学会会員。



林 良彦 (正会員)

1983年早稲田大学大学院理工学研究科博士前期課程修了。同年日本電信電話公社入社。現在 NTT 情報通信網研究所メッセージシステム研究部主任研究員。日英機械翻訳システム、日本文推敲支援システムなどの自然言語処理の研究に従事。1994年スタンフォード大学 CSLI 滞在研究員。人工知能学会、電子情報通信学会各会員。