自然言語処理 89－1
（1992. 5. 14）

文の連接パターンに基づく日本語テキスト構造の解析

田中智博　　林　良彦

ＮＴＴ情報通信網研究所

概要

談話やテキストでは、２文間に何らかの関係を持ちながら展開していることが多く、その連鎖の仕方には対象世界の特徴が反映されていることが多い。これに基づき、日本語テキストの構造を、隣接する２文の並びのパターン（連接パターン）を基に解析する手法を提案する。ここでは、単位文のタイプとそれを接続する接続属性の組み合わせにより、文をタイプ分類し、そのタイプの並びのパターンから単位文をノードとするネットワーク構造を導出する手法を述べる。また、クローズドデータ、オープンデータによる評価実験の結果から、本手法の有効性、解析規則の収束性について考察する。

A Japanese Text Analysis Method based on Sentence Adjacent Patterns

Tomohiro TANAKA　　Yoshihiko HAYASHI

NTT Network Information Systems Laboratories

Abstract

This paper proposes a new method to analyze cohesive relations between adjacent Japanese sentences based on constraints naturally contained in objective texts. We especialy 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 toward the closed data shows the effectiveness of our method. We also discuss the coverage of the rule set through an experiment toward open data.

1．はじめに

推敲支援、要約生成等に応用可能な日本語テキストの構造解析の検討を進めている。例えば、推敲を必要とするテキストでは、文脈の乱れ、不適切な表現の介入等が存在するので、入力される構造を予測することが困難であり、トップダウン的にあらかじめ用意した構造にあてはめていく（例えば[1]など）ことは難しい。また、テキストの構造を木構造として解析する手法（例えば[2]など）では、論旨展開が明確なテキストに対してその有効性が報告されているが、展開の乱れを含むようなテキストでは、その構造を十分に表しきれないと考えられる。そこで、それらの問題に対処する1方法として、テキストの構造を、隣接2文間の関係をベースとしてボトムアップ的に解析することによって、1つの述語表現とその格要素および副詞要素からなる単位文をノードとするネットワーク構造で表すことを目指している。

ここでは、ベースとなる隣接する2文間の関係を、テキストにより記述される世界の特徴を反映した文のタイプ分類及びその並びのパターン（以後「連接パターン」と呼ぶ）に基づいて解析する手法を提案する。まず、2章で本手法の基本的な考え方、処理の構成について述べる。次に、3章で個々の処理部について述べ、最後に4章で評価実験と考察について述べる。

2．テキスト構造解析

2．1　基本概念

テキストの構造には、結束性（cohesion）、首尾一貫性（coherence）が大きく係わっている[3][4]。一般に、多くの文章では、文が省略、照応、接続表現、語彙連鎖等の関係あるいは、密接な内容上のつながりを保ちつつ展開する[5][6]。本検討では、対象とするテキストを説明的文章とし、その典型的な例として応用上も有益であると考えられるソフトウェアのマニュアルをとりあげる。マニュアルのような不特定多数の読み手を対象とした説明的なテキストでは、書き手側の意志を読み手側に正確に伝えることが必要である[7]ために、特に、文間の内容上のつながりは強いと考えられる。また、機器の説明あるいは操作の説明といった比較的限られた世界での記述であるために、文のタイプが比較的限定される。そこで、この考え方に基づいて、テキストが記述する世界の特徴が文のタイプおよび文の連接に反映されると仮定する。この仮定から、文のタイプ分類を行ない、その連接パターンを基に隣接2文間の関係を解析することにより、テキストの構造解析の中心的部分を構成することができる。

2．2　処理概要

以上の考え方を基に検討を進めている文章構造解析の処理構成を図1に示す。入力は、テキストを構成する各文に対する構文解析結果の列であり、出力は、各文を構成する単位文をノードとし、それらの結束関係をリンクとするネットワーク構造である。ここでの結束関係とは、結束性の考え方に基づいて、文と文を接続する接続詞、接続助詞及び形式名詞等の接続表現を単位文単位に適用したものである。出力の概念例を図2に示す。このように単位文をノードとするネットワーク構造でテキスト構造を表現することにより、1文をノードとする木構造では表現しきれないテキスト構造も表現可能であり、不適切な文章構造の指摘などの応用へも適用可能となる。

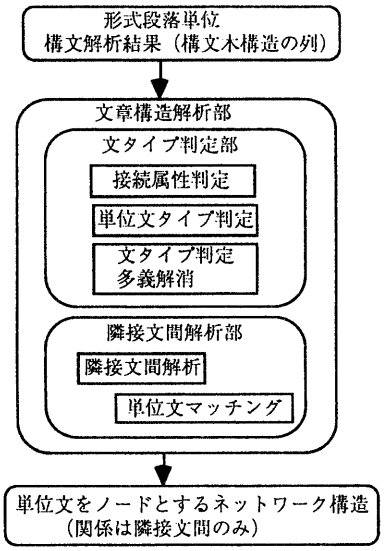


図1　処理の概要

3．隣接文間解析手法

3．1　文のタイプ分類

1）記述形式

単位文のタイプ分類を基に、文のタイプを分類する。実際のソフトウェアマニュアルでは、1文内で2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が多く（約60％）、このパターンを基本形とし、単位文のタイプとそれを接続する接続表現（以後「接続属性」と呼ぶ）の組み合わせにより以下のように記述する。

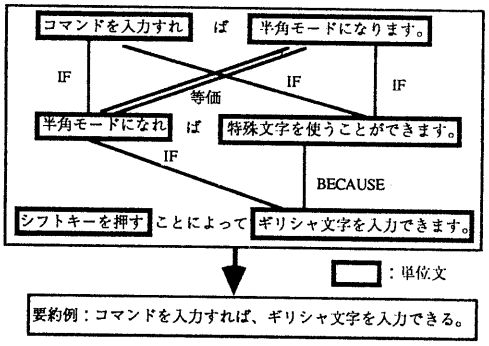


図２　構造概念例

表１　単位文タイプ分類 例

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ください）
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）

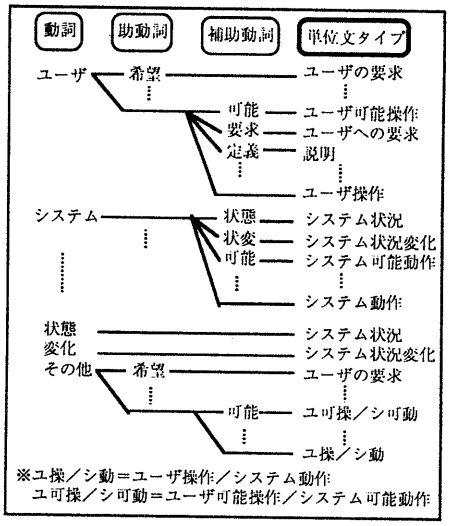


図３　単位文判定弁別ネットーワーク

（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３つ以上の単位文が存在する場合には、係り受け関係から上位にある２つで代表させる。

なお、接続属性の分類体系は、通常よく用いられている体系（例えば[8][9]など）に若干の修正を加えたものとした。

2）単位文のタイプ分類

文の主体、機能を主な観点として、単位文を分類した。対象としているソフトウェアのマニュアル文では、「（ユーザが）～すると、（システムが）～になる。」「（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、ユーザあるいは、説明対象のシステムである場合が多く、また、モダリティの部分に伝達上の機能（指示、状況説明、許容等）が現われる。主体は多くの場合省略されている[10]ので、ここでは主体として何をとりうるかという観点から動詞の分類をおこない、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。現在用いている体系の一部と弁別ネットワークの一部をそれぞれ表１、図３に示す。

3）単位文タイプ多義の解消

上で述べたようにタイプの決定において、主体の決定は、主に動詞の分類に依存している。しかしながら、動詞の分類だけではどうしても主体を一意に決定できない場合がある（例えば、「～を行なう。」がユーザの操作なのか、システムの動作なのか）。本検討では、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（例えば、「ながら」「つつ」で接続された２つの単位文は同動作主である。）を用いることにより、接続属性、単位文タイプ判定後の文タイプ内で絞り込みを行なう。

３．２　解析手法

対象世界の特徴は、文のタイプおよびその連接パターンに反映されると考える。従って、上記の文のタイプを基にその連接パターンに対して、可能な組

み合わせ、および結束関係を全て記述すれば、それを用いて隣接2文間の関係を解析することができる。一例を図4に示す。ここで2つの問題が考えられる。1つは、はたして、連接パターンだけで2文間の関係を一意に決定することができるのか、ということである。その解決策として、次の項で、関係の多義の絞り込みの一手法を提案する。また、もう1つは、実際に必要な全ての連接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。それについては、4章で、連接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。

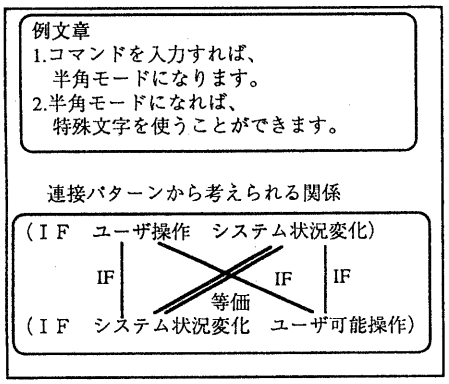


図4　連接パターン具体例

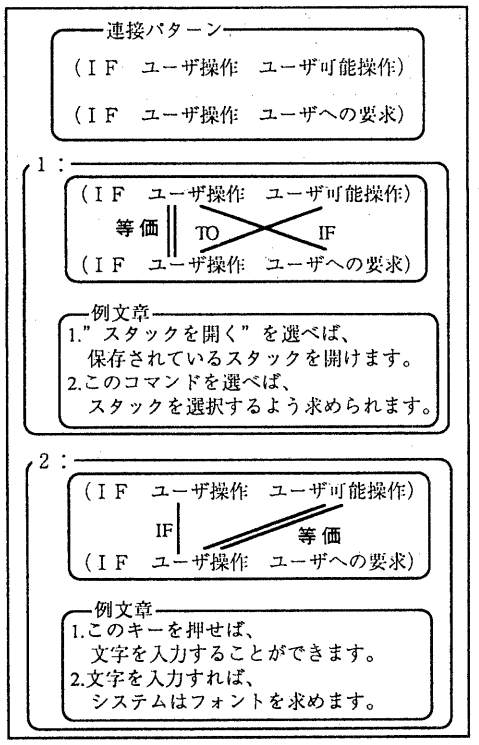


図5　連接パターン具体例

3．3　関係多義の絞り込み

図5に示すように連接パターンだけからでは可能な関係付けが一意に決まらない場合が存在する。しかしながら、図5の例に見られるように、前文の単位文と後文の単位文との関係の一部を、ダイナミックに決定することができれば、隣接文間の関係を一意に絞り込むことができる場合がある。この例では、等価性が成立する単位文の組み合わせ（図5で太字で示した部分）により切り分けが可能となる。このように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。そのアルゴリズムを図6に示す。2単位文を入力とし、各用言、各格要素のマッチングにより、単位文間に関係を付与する。なお、マッチングの判定の多くは、用言、格要素（名詞）、及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、対義のデータによっている。今後は、意味属性などの辞書情報の有効活用や、データの半自動生成などについても検討の必要がある。

3．4　解析規則

連接パターンをインデックスとした解析規則を図7のように作成し、これを基に解析を行う。解析規則は、まず、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、単位文間に結束関係を記述する（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、そのラベルとして結束関係名が付与される。ここで、付与する結束関係名は1文内における単位文間の接続関係および単文間の接続関係と同等であると考え、先に述べた接続属性に1文単位の関係（「そして」「また」「なお」等の関係）を加え、さらに、単位文マッチングで判定される等価、対比等の関係を加えたものとする。また、その規則内の全ての単位文マッチングが失敗した場合は、その連接パターン内で主となると考えられる前文の単位文と後文の単位文の間にリンク（ラベル：DEFAULT）をつける（γ）。

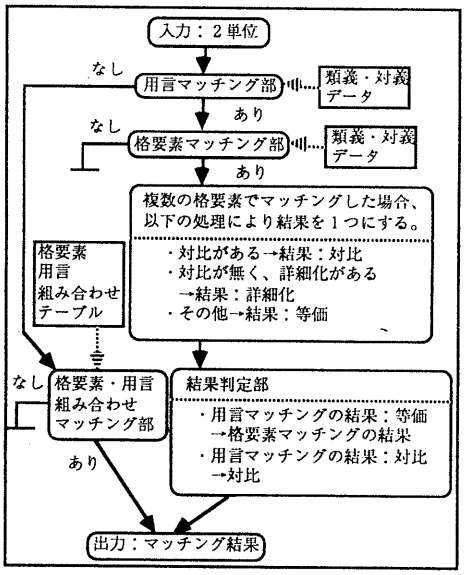


図6　単位文マッチングアルゴリズム

```
(def- cohesive-rule  rule_hc_if_1 ←― ルール名
   (:Chain_condition                          ┐
    (IF  ユーザ操作  ユーザ可能操作)          α
    (IF  ユーザ操作  ユーザへの要求)          ┘
   (:Syntax_condition                         ┐
    ((:Check_equal  _B  A)                    │
     (:Set_cohesive_link  _B  A  'equal)      │
     (:Set_cohesive_link  _A  A  'IF)         │
    ((:Check_equal  _A  A)                    │
     (:Set_cohesive_link  _A  A  'equal)      β
     (:Set_cohesive_link  _A  B  'IF)         │
     (:Set_cohesive_link  _B  A  'TO)         │
    ((:Check_contrast  _A  A)                 │
     (:Set_cohesive_link  _A  A  'contrast)   │
    (t                                        │
     (:Set_cohesive_link  _B  A  'DEFAULT) ……  γ
```

※ここで、A、Bは従文、主文を表し、
__は前文を表す。
:Check_equal：単位文マッチング（等価判定）
:Check_contrast：単位文マッチング（対比判定）
:Set_cohesive_link _A A 'EQUAL
：_AからAへEQUALの関係を付与する

図7　解析規則一例

4．評価実験・考察

4．1　評価実験

評価は、解析手法の精度および解析規則の収束性の両方を評価するために次に示す2つの実験により行った。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（解析規則が全ての連接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、机上検討により解析を行う文章の全ての連接パターンを抽出し、それ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文章と、実験を行う文章とを分けて実験を行う（オープンデータ）ものである。

机上検討

各実験では、実験に使用したマニュアル文章（オープンデータにおける規則作成用文章は除く）に対して、あらかじめ机上検討により、文タイプの判定、結束関係の付与を行う。その際、以下に示すような、①隣接文間に何らかのつながりがあるにもかかわらず結束関係が明確でない（単位文単位に明確な関係が付与できない）場合、②隣接2文間を越えて結束関係が成立する場合は、本手法の対象外としてあらかじめ除外しておく。図8にマニュアル約1500文に対して調査したこれらの割合を示す。

①例：
「カーソルが16バイト目を過ぎると、次の16バイトが表示されます。
リターンキーを押すと、元に戻ります。」
（文単位で継続「そして」の関係が成立している。）
②例：
「日付を変更する場合は、”DAY”を入力してください。
日付を半角文字で入力します。
日付を変更しない場合は、そのままリターンキーを押してください。」
（第3文が第1文と関係している。）

前者の場合は、文単位で関係付けが行われているために、単位文単位で結束関係を付与できないのが原因である。それらの内訳としては、動作の継続、補足説明等の接続関係によって関係付けられている（約90％）、省略、照応、語彙の連鎖等の関係のみによって関係付けられている（約10％）である。今後、文単位の関係（例えば接続詞によるものなど）と、結束関係との係わりについて検討していく必要がある。

後者の場合の対処については、隣接文間解析を補足するものとして、接続詞等の表現に着目した解析手法の手がかりを5章で述べる。前者の場合同様、

ここでも２文間の関係が文単位になることが多く、上記の検討が必要となる。
次に各実験における結果を述べる。

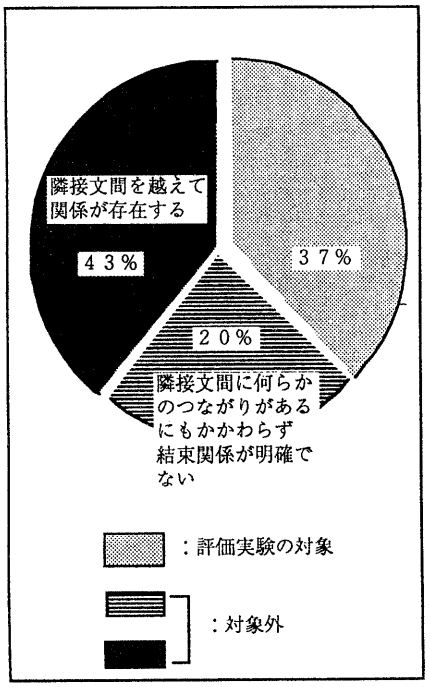


図８　マニュアル調査結果（対象約１５００文）

クローズドデータ

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段落から解析規則を作成した。規則数は、236で、このうち文タイプの連接パターンによる制約のみから結束関係が定まる規則数は、141であった。次に評価指標を示す。

文タイプの判定

机上検討によって得られた文のタイプと一致した割合（**文タイプ判定率**）と、文タイプ内で多義が生じたものに対して、正しく解消できた割合（**文タイプ多義解消率**）とする。

文タイプ判定率＝

$$\frac{\text{机上検討によって得られた文のタイプと一致した文数}}{\text{実験に用いた文の総数}}$$

文タイプ多義解消率＝

$$\frac{\text{多義が正しく解消できた文数}}{\text{多義が生じた文の総数}}$$

隣接文間解析

机上検討によって得られた関係と一致した割合（**規則適合率**）とする。

規則適合率＝

$$\frac{\text{机上検討により得られた関係と一致した連接パターン数}}{\text{解析規則により関係付けた連接パターン数}}$$

また、単位文マッチングの効果をみるためにこれをすべてスキップする実験も合わせて行った。

結果：文タイプ判定率は約80％、文タイプ多義解消率は約87％、単位文マッチングを行った場合の規則適合率は約81％、単位文マッチングを行わずに１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、クローズドデータに対する有効性、単位文マッチングの効果を確認することができた。

オープンデータ

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、同連接パターンの出現頻度をパラメータとして、解析規則を作成した。表２に各パラメータに対する規則数を示す。なお、出現頻度は、不要な規則の出現を抑制するため頻度２以上を対象とし、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。次に評価指標を示す。

表２　解析規則作成結果

出現頻度＼作成用文数	471文	1104文	1548文
５以上	１２	３７	５２
３以上	３４	７７	８８
２以上	６８	１３６	１６６

（解析規則数）

隣接文間解析

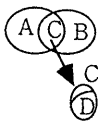
クローズドデータで用いた**規則適合率**と、作成された規則と実際に使われた規則との割合（**規則使用率**）、規則によって関係付けられた連接パターンの全体に対する割合（**規則適用率**）の３つとした。

規則適合率＝

$$\frac{\text{机上検討により得られた関係と一致した連接パターン数 (D)}}{\text{解析規則により関係付けた連接パターン数(C)}}$$

規則使用率＝

$$\frac{\text{解析に使用された解析規則数 (C)}}{\text{作成された解析規則の総数 (B)}}$$



規則適用率＝

$$\frac{\text{解析規則により関係付けた連接パターン数 (C)}}{\text{机上検討により関係付けた連接パターンの総数 (A)}}$$

結果：各結果を表3に示す。これによると、規則適合率は約88％で、オープンデータに対しても本手法の有効性が確認できた。また、規則適用率は、作成用文数約1500文、同連接パターン出現頻度２以上で、約64％を得た。規則使用率は、同連接パターンの出現頻度によって規則数を絞り込むことにより増加している。このことから、作成用文数を増やすとともに、同連接パターンの出現頻度によって規則数を絞り込むことにより、有効な規則を作成することができることを確認した。

表3　オープンデータによる評価実験結果

A．規則適合率（%）

出現頻度＼作成用文数	471文	1104文	1548文
2以上	86	88	88

B．規則使用率（%）

出現頻度＼作成用文数	471文	1104文	1548文
5以上	92	68	62
3以上	68	49	44
2以上	47	36	33

C．規則適用率（%）

出現頻度＼作成用文数	471文	1104文	1548文
5以上	22	42	47
3以上	38	54	54
2以上	45	60	64

4．2考察

以下に、各処理における失敗の原因とその対策を示す。また、最後に解析規則の収束性について考察する。

文タイプ判定

失敗の原因としては、単語切り等の形態素レベルでのミスを除けば、「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が挙げられる。これらの多義は、文解析の段階で精度良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、全て単文である。接続属性、他の単位文の情報が欠落していることが失敗の原因である。これらの情報以外で判定する方法を検討する必要がある。

隣接文間解析

クローズドデータ、オープンデータとも連接パターンに単文が含まれる場合、文タイプ中の接続属性による情報が欠落するために、解析規則作成時に関係が絞り込めていない。これを原因とする失敗が約半数を占める。

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。

単位文マッチング

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は、文脈によらず等価と考えられる。よって、類義関係として定義しておけば表層上から判定が可能であり、本研究で提案した単位文マッチング手法が有効である。しかし、例２に示すような表現の等価性が文脈に依存するものや、例３に示すような現在のマッチングでは関係が扱え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[3]についても検討していく必要がある。

解析規則の収束性

オープンデータによる実験結果から得られる規則作成用文数と規則適用率、作成された規則数の関係を図９に示す。今回の実験条件では、解析を行った文の約４倍の数の作成用文により解析ルールを作成した結果、約64％の適用が得られた。しかしながら、図９に見られるように、作成用文数を増加させるだけでは、可能な全ての関係を記述しつくすことはかなり困難である。今後は、解析規則の抽象化、効率的な作成方法の検討により、解析規則を自動作成する手法を検討する必要がある。

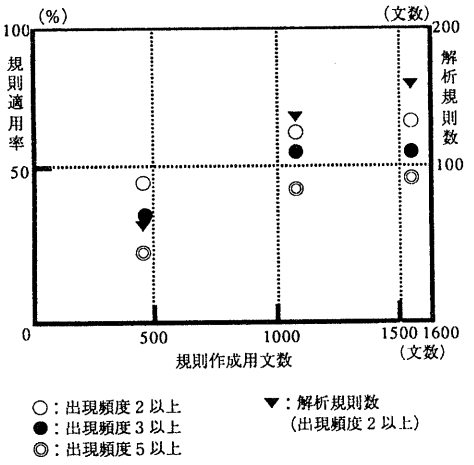


図９　作成用文数と適用率、規則数の関係

５．むすび

テキストは表層上あるいは、内容的に密接なつながりをもって展開する。本論文では、この考えに基づき、対象世界を反映する文のタイプ分類と、その連接パターンから、テキスト構造解析のベースとなる隣接２文間解析手法を提案し、その有効性について述べた。今後は、解析規則の抽象化、単位文マッチングの精度向上を図ると共に、隣接２文間解析を補足するものとして、隣接文間では扱えない離れた文間の関係を、次に示すような表現に基づいて解析する手法を確立していく。現在検討してる手がかりとなる表現としては、次のようなものがある。

①接続詞及び相当表現

②特定の文タイプの組み合わせ

例：

「日付を変更する時、～してください。

日付を変更しない時、～してください。」

③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、～キーを押してください。

日付は、～コマンドで見ることができます。」

これらを机上検討によって調査した結果、上記の手がかりを基に、隣接２文間解析で扱えなかった表現の約70％を補完可能であることが分かった。しかしながら、個々の処理において、接続詞の機能の検討、接続先の同定、基本となる単位（文、単位文等）、付与する関係等の問題、また、上記手がかりは共起することが多いので、規則適用の制御方法等の問題がある。

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 (1989)

[2] 小野、浮田、天野：文脈構造の分析、自然言語処理研究会 70-2 (1989)

[3] Hobbs.J：Coherence and Coreference , Cognitive Science Vol.3 No.1 (1979)

[4] Halliday.M.A.K, Hasan.R：Cohesion in English, Longman London and New York (1976)

[5] Jane.M, Graeme.H ：Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text , Computational Linguistics Vol.17 No.1 (1991)

[6] 市川：国語教育のための文章論概説、教育出版 (1978)

[7] Brad M.McGehee（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、日経マグロウヒル社 (1987)

[8] 寺村：日本語の文法（下）、国立国語研究所 (1981)

[9] 南：現代日本語の構造、大修館書店 (1974)

[10] 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、情処論 Vol.31 No.10 (1990)

[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 (1992)