

## 対話的文書検索のための文書構造解析

住田一男 小野顕司 三池誠司

(株)東芝 研究開発センター

対話的な文書検索のための自動的な文書構造抽出システムを開発した。章や節などの階層関係と、各章や節を構成する文章本体内部の修辭的な構造を抽出する。この修辭的な構造は、文と文の間の関係や文のまとまりの間の関係を、接続詞や文末の表現を手掛かりとして解析する。抽出される構造に基づいた検索機能、検索結果の効率的な提示のための自動抄録機能を実現した。検索結果の提示という観点から、自動抄録機能に関して評価実験を行い、有効性を確認した。

## DOCUMENT STRUCTURE EXTRACTION FOR INTERACTIVE DOCUMENT RETRIEVAL

Kazuo Sumita Kenji Ono Seiji Miike

Research and Development Center, Toshiba Corp.

Komukai-Toshiba-cho, Saiwai-ku, Kawasaki 210, Japan

We have developed an automatic document structure extraction system for interactive document retrieval. The system extracts a section organization representing a hierarchy of chapters and sections in the original document, and extracts a rhetorical structure for the body of each chapter or section. The rhetorical structure represents relations between sentences and those between the chunks of the sentences, which are extracted based on linguistic clues such as connectives, sentence predicates. On the basis of the extracted structures, a document retrieval function and an automatic abstract generation function are realized. From the point of view of document presentation, the abstract generation function was evaluated.

## 1 まえがき

対話的文書検索を実現するために必要となる文書の構造化について考察する。なお、検索対象としては、文献抄録のような短い文書ではなく、全文からなる文書を想定している。

近年、ワークステーションの計算機パワーの増大にともない、全文文書を検索対象とした全文検索システムの実用化が進みつつある。しかし、現在実用化されている全文検索システムでは、検索命令に含まれる語句が文書中に含まれるか否かを判定する機能が実現されているにすぎない。検索時に、検索語句がどのような内容を述べている部分で用いられているかなどという点については考慮されていない。

研究論文のような比較的長い文書は、「まえがき」や「あとがき」など複数の章や節を含んでいる。しかも、「まえがき」には「技術的背景」や「研究目的」などの内容が述べられている一方、「あとがき」には「得られた結論」や「今後の課題」などの内容が述べられている。背景が述べられている部分で用いられている語句や、結論が述べられている部分で用いられている語句を区別して検索することで、より高精度な検索を実現できる可能性がある。

また、検索するという行為は、単に文書を検索して終りというものではない。むしろ、検索した文書を読み、理解することにより、内容を参考にしたり、再利用するということが本来の目的であろう。このような点から、検索行為の効率化を支援するためには、検索結果の提示についても配慮が必要である。

例えば、複数の文書が検索された場合、文献抄録のように検索単位が十分短ければ、全文を提示しても検索者の負担にはならないかもしれない。しかし、長い文章を検索単位とする場合、全文を提示するだけでは、個々の文書の内容を理解するために読んでいくことが検索者の負担となる。効率的な検索結果の提示を実現するためにも、文書の何らかの構造化が必要となってくる。

ハイパーテキストは、このような観点に立った対話的な文書提示の一つのアプローチと言える。しかし、構造化を人手で行わなければならないという点が、多量の文書を処理対象とする全文検索システムに適用す

る上で問題となる。

従来の検索システムでは、検索対象の中心が文献抄録などのように比較的短い文書を検索対象とする場合が多かった。短い検索対象の場合、文書の構造を認識する必要性は小さい。しかし、今後全文文書の増大にともない、比較的長い文書を検索対象とする場合が多くなっていくと予想される。このような全文を対象とした検索において、精度の高い検索や効率的な文書提示を実現するためには、文書の自動的な構造化が必須の要素技術となってくる。

本稿では、効率的な文書検索のための文書構造解析手法について検討する。この構造解析の目的は、検索処理における精度の向上と効率的な文書提示の実現とにある。そこで、これらの2面から文書の構造化を検討する。

検討した結果に基づいて文書構造解析システムを、ワークステーション上で試作した。また、検索結果の文書を提示する機能として、文書構造解析に基づいた抄録生成システムを試作した。この抄録生成システムを、文書提示の観点から評価を行ったので、その結果についても述べる。

## 2 関連研究

従来、自動的な文書構造の分析に関する研究がいくつか行われてきた。情報検索以外の文書処理の分野で、自動的な文書レイアウトシステムのために文書構造が検討されている [Iwai 89]。しかし、章や節の階層関係の抽出にすぎず、文書中の意味的な役割を認識することは、対象外であった。また、自然言語処理の分野で、文の間の論理的な関係を取り扱うために、言語的な手掛かりに基づいて文脈構造の分析が検討されている (例えば、[Cohen 87], [Grosz 86], [Mann 87], [福本 92], [Sumita 92])。しかし、一般的な観点から構造化が議論されており、具体的なアプリケーションへの適用が考慮されていない。

また、技術論文を対象として、「目的」や「結果」などを表現している文を、構文パターンを用いて抽出する試みが行われている [矢島 91]。しかし、文レベルの処理にとどまっているため、複数の文にわたって表現されている場合、抽出できないことがある。そこ

で、接続表現などの情報を手掛かりに複数文で表現される場合に対応する方法が検討されている [西村 93]。しかし、局所的な処理しか行っておらず、構造上のまとまりを認定する上で問題があると考えられる。

情報検索の分野でも、検索精度を向上させるため、文書構造や文脈構造の利用が検討されつつある。(例えば、[Hearst 93])。[Liddy 92] では、ニュース文のような長さの短い文章を対象にした構造化が検討されている。構造の表現は、主要な出来事や関連する出来事などのサブトピックスの単純なリストであり、効果的な文書提示のためには必要と考えられる階層的な構造が表現できない。

[Burkowski 92] は、SGMLのような構造化文書に対して、検索モデルの形式化を行っている。あらかじめ、構造化された文書を対象としたものであり、構造の抽出に関する検討はされていない。

### 3 対話的検索のための文書の構造化

#### 3.1 文書の構造化

全文からなる文書の効率的な検索を実現するため、文書の構造化について考察する。

科学技術論文など比較的長い文書は、通常複数の章や節からなっている。章や節は、その文章の書き手が持っている意味的なまとまりに対応しており、検索処理や提示処理においても区別して扱える必要がある。

近年、SGMLやODAなどの標準仕様や、TEXに基づいて、定義したタグを用い構造を定義した文書が増えつつある。このような文書では、章や節の構造は明確に定義されており、改めて解析する必要はない。しかし、今だに大半の文書が、ワープロなどで作成されたフラットな文書であるのが現実である。特に、全文検索システムでは、OCRによって入力した文書を検索対象とする場合も多い。このような場合についても、文書構造の明示的なタグは付与されていないため、構造解析の処理が必要となってくる。

章や節は、その見出しとともにその本体であるいくつかの文からなる文章を含んでいる。これらの文の論理的あるいは意味的な関係を捕らえるためには、これらの文章の構造についても解析する必要がある。

本稿では、検索処理の精度向上と効率的な文書提示

の実現とを目的としている。検索精度の向上のため、文の意味的な役割を抽出する。また、効率的な文書提示のため、検索結果の抄録を自動的に生成する。これらの機能を実現するため、文章の構造解析では、文間の関係を捕らえるだけではなく、内容上のまとまりなどの認識も必要と考えられる。

ここでは、2つの文書階層（文書構造と文脈構造）から文書の構造を抽出する。文書構造は、文書の章や節の階層構造を表現する。一方、文脈構造は、各章や節内の文や文のまとまりの間の論理的な関係を表現する。

文書構造については、章や節の単位やその階層的な関係は客観性が高く、表現上の問題はさほど生じない。また、その分析についても、章見出しや章番号、インデントーションなどを手掛かりとして分析することができる。しかし、文脈構造については、どのような言語単位を解析の対象とするのか、それらの言語単位をどのような関係で関連づけるのか、というような表現上の問題が大きい。

文脈構造の解析あるいは表現単位としては、段落、文、クローズ(1文内の主節や従属節)などが候補として考えられる。「背景」や「関連研究」のようなサブトピックスの抽出を考えた場合は、段落単位の解析で十分かもしれない。しかし、文書検索結果の提示機能として想定している自動抄録では、段落は表現単位としては大き過ぎる。

本稿では、文脈構造の表現の単位を文とし、抄録生成についても文単位で取捨選択を行うものとする。文は、意味的なまとまりとして明確に区切られた単位である。したがって、クローズ単位の処理に拡張する場合も、文を基本単位とする処理に追加する形で対応できると考えられる。

文脈構造の抽出は、接続詞や照応表現、文末表現など言語的な手掛かりに基づいて行うものとする。解析対象の文章の話題に関連する世界知識がなければ、正確な構造が得られない場合もある。しかし、あらゆる分野の世界知識を準備することは困難であり、検索システムを前提とした処理では、分野を限定することも現実的ではない。そこで、ここでは、世界知識には頼らない方法とする。

### 3.2 文書構造解析

ある技術論文の構造を、文脈自由規則で表現すると図1のように表現できる。

文書	→	序, 節集合, 参考文献.
序	→	タイトル, 著者, 所属, 概要.
節集合	→	節, 節集合.
節	→	見出し, 文章本体.
概要	→	見出し, 文章本体.
...		

図 1: 文書構造の表現

文書が、序と複数の節と参考文献からなり、各節が、見出しと文章本体 (複数の文からなる文章) からなることを表している。

文書自動レイアウトのため、文書構造の解析が提案されている [Iwai 89]。本稿における文書構造解析においても、ほぼ同様の処理を行う。

具体的には、見出しを認識するために、先頭が数字や記号で始まるという情報、行の末尾に句点が存在していないという情報、などの情報を文章中より取り出している。また、ある行の先頭が空白であることを検出し、その位置が段落の開始であることを認識する。

文書構造の解析では、節ごとに見出しとその節を含む文章本体の範囲、その文章本体を構成する各段落の先頭の位置、などを出力する。そして、この解析結果は、原文の文書と対応してデータベースに格納する。また、各章の文章本体は、次の文脈構造解析においてその内部構造を解析する。

### 3.3 文脈構造解析

#### 3.3.1 文脈構造の表現

本稿では、文脈構造の表現を、自然言語文の統語構造のアナロジから2分木で表現している。統語構造中の任意の部分木が、文法的な構成要素をなすように、文脈構造における部分木は議論上の一構成要素をなす。ある部分木は別の部分木を直接の子ノードとする木に付与される関係によって範疇化される。

そして、文脈構造は、段落間の構造と段落内の構造の2階層からなるものとしている。すなわち、段落間の構造は段落を終端ノードとする構造を、段落内の構造は文を終端ノードとする構造をそれぞれ表現している。

図2の文章例 (下線部は文脈構造解析の手掛かりとなる表現である) に対応する文脈構造を図3に示す。

- 1: 区間分割相関方式は、収束が保証されているという利点をもっている反面、全区間をひとわたり修正するのに、トレーニング信号をN回受信しなければならず、収束に時間がかかるという欠点をもつ。
- 2: そこで、収束を速くするための便法として、区間の仕切を取り払い、トレーニング信号受信のたびに(21)式を全区間 [L, M] に一斉に適用する方式を考えた。
- 3: これを、“部分相関方式”と呼ぶ。
- 4: 部分相関方式では、当該サブ区間以外のタップ利得が一時的に凍結されているという条件は成り立っていないので、収束は理論的には保証されていない。
- 5: しかし、修正係数  $\alpha$  を十分小さく選べば、近似的にこの条件が満足されているから、実用上は収束が期待しうる。

図 2: 文章例

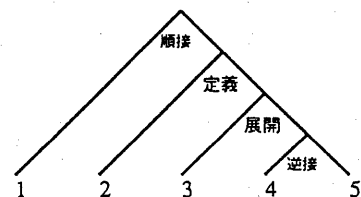


図 3: 文脈構造

図3の文脈構造において、数字は図2で示した文章中の各文の文番号を表している。また、“逆接”、“順接”

などは文間の関係を表しており、本稿では修辞関係と呼ぶ。

図3に示す文脈構造は、第5文が第4文に対して逆接という関係に、第3文から第5文のまとまりが第2文に対して定義という関係に、そして、第2文からと第5文のまとまりが第1文に対して順接という関係にあることを示している。

修辞関係は、接続詞や文末表現など文間の接続的な関係を明示する表現を、34種類に分類したものである。表1に修辞関係の例と、その修辞関係に相当する表層的な表現例を示す。

表 1: 修辞関係の例

関係名	表現例
順接	したがって、そうすると...
理由	なぜなら、というのも、...だからである。
対比	一方、他方
例示	例えば、この他、...は一例である。
背景	従来、...されるようになってきた。
逆接	しかし、それにもかかわらず
話題	本特集では、ここでは
定義	これを...と呼ぶ
展開	この、その

次に文脈構造解析では、この修辞関係を、文と文の間の関係だけを表現する関係子としてではなく、文のまとまりの間の関係を表現する関係子として取り扱う。

### 3.3.2 文脈構造の解析

文脈構造の解析では、文書構造で得られた各節の文章本体に対して、以下のような手順で解析を行う。

#### (1) 文解析

まず始めの処理として、文章本体中の各文に対し形態素解析ならびに構文解析を行う。

#### (2) 修辞関係の抽出

このステップでは、各文の形態素結果あるいは構文解析結果に対して、その文に先行する文脈に対する修辞

関係を取り出す。そして、各文の文番号と修辞関係からなる系列を出力する。この系列を修辞関係系列と呼ぶ。

表1で例示した表現パターンと修辞関係の対応情報は、システム内ではテーブルの形で表現している。このテーブルには、表層表現、形態素解析列までのパターンが正規表現で記述できる。そして、最長一致するパターンに対応する修辞関係が、その文の修辞関係として選ばれる。

#### (3) セグメンテーション処理

ステップ2で解析される修辞関係系列内の各修辞関係は、解析対象の文ごとに求められた関係にすぎない。一方、文章中には、複数の文にわたって構造を規定するような修辭的な表現も存在する。セグメンテーション処理では、このような複数文わたる表現から、構造化に対する制約情報を取り出す。制約情報はセグメンテーション規則と呼ぶ規則に記述し、ルールベースで処理を行う。

図4に、セグメンテーション規則の例を示す。図示した規則は、“確かに”や“本来なら”という表現が文頭にくる文の何文か後に、逆接の修辞関係の文が来た場合、その文の直前でまとまるという知識を記述している。

入力例	確かに...。...。しかし...
出力	[ 1... ] @ 逆接 2
表層	(確かに   本来なら) -
関係	- 逆接

図 4: セグメンテーション規則の一例

#### (4) 構造候補生成

セグメンテーション処理で付与された制約を破らない2分木構造を生成する。そして、生成された構造のうち、得点の高い部分構造を持つ構造を優先する。この得点づけは、文脈構造上隣接している修辞関係に基づいて決定する。

段落間の構造は、段落の先頭の文の修辞関係に基づいて、段落内の構造生成と同様に解析する。

文脈構造も、文書構造と同様に原文の文書と対応さ

せてデータベースに格納する。

#### 4 意味役割の抽出

文書中には、例えば、技術論文では「目的」や「背景」、「結論」というように典型的な意味的役割を持つ部分が存在する。前節で説明した文書構造と文脈構造に基づいて、このような意味役割を判定し、文書中の位置を抽出する。

抽出すべき意味役割は、分析対象の文書カテゴリによって異なる。例えば、技術論文では「目的」や「背景」、「結論」などが典型的なものであるが、また解説文では「背景」や「話題」、社説では「背景」や「意見」、「結論」などが相当すると考えられる。

処理は、ルールベースで行う。部分的な文脈構造を左辺とし、意味役割とその役割に対応する文章セグメントを右辺とする規則（意味役割抽出規則）に基づいて、各意味役割に対応する文章セグメント（へのポイント）のペアを抽出する。意味役割抽出規則の一例を図5に示す。

```
TOD ; (int#1, 話題, sub#2) ; sub#2 ; EOC ;
      int#1 ; int#1 ; ; 話題.
EOD ; sub#1 ; sub#1 ; EOC ;
      (int#1, 意見, 2) ; 2 ; ; 結論.
```

図 5: 意味役割抽出規則

図5において、“TOD”や“EOD”は、対象となる文が現れる節の文書中の位置(“TOD”は最初の節、“EOD”は最後の節)を表している。また、“TOC”や“EOC”は、対象となる文が現れる段落の節中の位置(“TOC”は最初の段落、“EOC”は最後の段落)を表している。(int#1, 話題, sub#2)や sub#2などは、照合する文脈構造の部分構造を表現しており、その中で、int#1は非終端ノードを、sub#1は段落レベルの終端ノードを、数字“2”は文レベルの終端ノードをそれぞれ表現している。

例えば、ある文書中の最初の節に次に示すような段落レベルの文脈構造が存在する場合、最初の規則が照合する。

((sub#1, 展開, sub#2), 話題, sub#3)

段落 sub#3 に対応する文番号が、文役割“話題”に相当する部分として取り出される。

#### 5 抄録生成

文書構造と文脈構造に基づいて抄録文を生成する[知野 93]。利用者の指定に従って、タイトルの抜粋だけから、特定の章についての変長の抄録を生成することができる。

各章の抄録生成では、構造化した文脈構造を参照し、あらかじめ定めた修辭関係の前後ノードの相対的な重要度に基づいて抄録文を生成する。修辭関係の相対的重要度は、3つのカテゴリに分類している。一例を表2に図示する。

表 2: 修辭関係の相対的重要性

タイプ	修辭関係 (表現例)	重要なノード
<i>RightNucleus</i>	順接 (よって), 逆接 (しかし), ...	右
<i>LeftNucleus</i>	例示 (例えば), 補足 (ただし), ...	左
<i>BothNucleus</i>	並列 (また), 対比 (一方), ...	等価

例えば、*RightNucleus*に分類された順接は、右ノードが左ノードに比べ重要であることを意味している。

各ノードの重要度を決定するために、修辭関係の相対的重要度に応じて左右ノードにペナルティを与え、ペナルティの大きいノードを再帰的に薊りとり、抄録文を生成する。*RightNucleus*に分類された関係では、左ノードに1ペナルティを、また、*LeftNucleus*に分類された関係では、右ノードに1ペナルティを、与える。トップノードから順次このペナルティを加えることにより、各中間ノードについてのペナルティを計算することができる。このペナルティの大きいノードから薊りとる。最終的な構造中の終端ノードを抄録文として選択する。なお、抄録文として選ばれた文数が

規定の文数以下に達しない場合は、前方の文を優先して残すものとする。

最終的な抄録生成においては、接続表現の付け換えも行っている。

図3の構造に対して、上記のペナルティ計算を行うと、図6のようになる。破線はペナルティの境界を表しており、例えば、第1文、第3文、第5文には1ペナルティが、第4文には2ペナルティがそれぞれ課せられている。

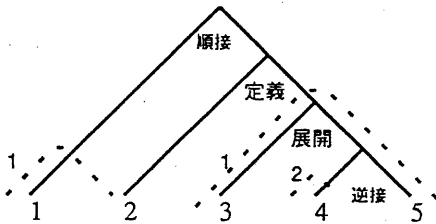


図 6: ペナルティ計算

この場合、最短の抄録は、第2文目だけとなる。また、次に長い抄録は、第1文目と第2文目の2文からなる抄録である。

## 6 インプリメンテーション

中本らが開発した全文検索システム [中本 93] の全文検索機能を利用し、文書構造に基づいた検索処理と、抄録生成のインプリメントを、ワークステーション (東芝 AS4000 シリーズ) を用いて進めている。

文書の構造解析はかなり処理時間が必要となるため、文書検索時に行うのは、現状の計算機パワーからすると現実的ではない。そこで、文書の構造解析と意味役割の抽出処理に関しては、実際の文書検索に先立って行う形を取っている。文書構造や文脈構造、文の意味役割の抽出結果は、テキストデータベース中の各文書と対応させてデータベース中にあらかじめ格納しておくようにしている。

一方、抄録生成は、処理パワーをさほど必要としないので、これらの格納情報を参照し、検索された文書ごとに検索時に実時間で処理するようにしている。これにより、詳細度の異なる抄録を動的に表示すること

ができる。また、文書構造に基づく検索処理に関しても、意味役割などの格納情報に基づいて実時間で処理するようにしている。この検索処理の詳細については、[三池 93] で詳述する。

## 7 抄録生成に関する実験

開発した抄録生成部分について文書検索における文書提示としての評価を行った。その結果を述べる。

評価の方法は以下の通りである。

- 問題作成者が原文の文書集合を参考に設問を作成する。
- 問題作成者とは異なる被験者に原文の文書集合あるいは抄録文の文書集合を与える。
- 被験者はすべての設問に対して、設問に対応する文書をすべて列挙する。
- 設問ごとに被験者が見つけられた文書数と正解の文書数の比 (再現率) を求める。

なお、一人の被験者に同じ内容の原文と抄録文とを提示することはできないので、別々の被験者が行った結果を比較することになる。

また、今回の実験では、実験のパラメータを以下のように設定した。

- 一度に与えられる原文もしくは抄録文の文書数を20文書とした。
- システムが生成する抄録文を原文との文数比が30%になるように設定した。
- 設問の数を7問とした。
- 問題を解く時間を一定 (5分) とした。

評価結果を表3に示す。表3は、与えられた設問に対する再現率について、抄録文提示の場合と原文提示の場合とで比較した表である。

再現率は、全設問について正しい文書を列挙できた場合が1となる。したがって、再現率が高いほど文書提示として効果があることになる。

表3によれば、平均的に見て、抄録文提示の方が原文提示に比較して再現率が高い。

表 3: 抄録文と原文の比較 (A, B, C は被験者)

文書集合	集合 1	集合 2	集合 3	平均
抄録	0.92 (A)	0.71 (B)	0.29 (C)	0.64
原文	0.28 (C)	0.50 (A)	0.43 (B)	0.40

文書集合 3 については、原文提示の方が再現率が高い。しかし、抄録文提示について評価した被験者 C は、原文提示と抄録文提示との両方に関して、他の被験者に比べ再現率が低い。このことから、文書集合 3 における再現率の逆転は、被験者間の差によるものと考えられる。

評価データの量や被験者数が少なく、プリントアウトした評価データに基づいた評価である。客観的な結論を下すには、さらに多くの被験者で実験する必要がある。しかし、この評価結果は、抄録の提示が文書検索の提示機能として有効であることを示唆していると考えられる。

## 8 まとめ

対話的文書検索システムにおいて効率的な検索を実現するため、文書の構造抽出システムを開発した。抽出する文書の構造に基づき、任意の長さでの抄録文の自動的な抄録生成や、文の意味役割の抽出を行う機能を実現した。これらの機能は、効率的な検索結果の提示と高精度な検索を行うために必須の機能である。

構造解析の情報を利用した検索については、別途実験およびインプリメントを進めている。その内容については、[三池 93] で詳述しているので参照されたい。

本稿で述べた抄録生成に関する評価実験は、紙の上にプリントアウトしたものをを用いた。検索システムの表示手段とした場合の評価や、検索作業全体に関する評価などは、まだ行っていない。今後の課題である。

自然言語は、知識を表現したり伝達する上で最も柔軟性に富んだ媒体である。今後、情報検索が、計算機の利用形態として最も重要な位置を占めていくと考えられる。その中で、知識表現として自然言語の重要性

は増しこそすれ、低くなることはない。自然言語の特長を有効に利用して、情報検索の高度化を図ってきたい。

## 謝辞

本研究を進める上で、有益な議論をして頂いた(株)東芝 研究開発センターの竹林洋一、平川秀樹、武田公人、伊藤悦雄、水谷由美諸氏に深謝するとともに、全文検索システムを快く貸与して頂いた(株)東芝 情報処理・機器技術研究所の岩井勇課長に深謝いたします。

## 参考文献

- [Burkowski 92] Burkowski, F.J.: "Retrieval Activities in a Database Consisting of Heterogeneous Collections of Structured Text", *Proc. 15th Annual Int. SIGIR'92*, pp.112-125, 1992.
- [Cohen 87] Cohen, R.: "Analyzing the Structure of Argumentative Discourse", *Computational Linguistics*, Vol.13, pp.11-24, 1987.
- [Grosz 86] Grosz, et al.: "Attention, Intentions and the Structure of Discourse", *Computational Linguistics*, Vol.12, pp.175-204, 1986.
- [Hearst 93] Hearst, M.A.: "Cases As Structured Indices for Full-Length Documents", *AIII'93, Spring symposium series, CBR/IR: Exploring Opportunities for Technology Sharing*, pp.140, 1993.
- [Iwai 89] Iwai, I., et al.: "A Document Layout System Using Automatic Document Architecture Extraction", *Proc of CHI'89*, pp.369-374, 1989.
- [Liddy 92] Liddy, E.D., et al.: "DR-LINK Project Description", *SIGIR Forum*, Vol.26, No.2, pp.39-43, 1992.
- [Mann 87] Mann, W.C. and Thompson, S.A.: "Rhetorical Structure Theory: A Framework for the Analysis of Texts", *USC/Information Science Institute Research Report RR-87-190*, 1987.
- [Sumita 92] Sumita, K., et al.: "A Discourse Structure Analyzer for Japanese Text", *Proc. Int. Conf. Fifth Generation Computer Systems 1992 (FGCS'92)*, pp.1133-1140, 1992.
- [知野 93] 知野哲朗他: "日本語論説文自動抄録システムの試作と評価", 第 46 回情処全大, 第 3 分冊, 1993.
- [中本 93] 中本幸夫他: "日本語解析を用いたフルテキストサーチの実験", 第 46 回情処全大, 第 3 分冊, 1993.
- [西村 93] 西村健士他: "特定表現の重点的解析による科学技術論文構造化手法", 情処研資, FI-29-5, 1993.
- [福本 92] 福本淳一他: "文の接続関係解析に基づく文章構造解析", 情処研資, NI-88-2, pp.9-16, 1992.
- [三池 93] 三池誠司他: "文書の構造解析に基づく文書情報検索", 情処研資, FI-31, 1993.
- [矢島 91] 矢島真人他: "文書への意味属性付与のための意味辞書の開発", 第 43 回情処全大, 第 3 分冊, 1991.