ICOT Technical Report: TR-325

TR-325

内容検索のための自然言語パーサ

伊吹　潤、杉山健司、鈴木香緒里、玉田郁子
川崎正博

November, 1987

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03) 456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

内容検索のための自然言語パーサ

Natural Language Parser for Content Retrieval

伊吹　潤，杉山健司，鈴木香緒里，玉田郁子，川崎正博

（富士通）

我々はユーザの自然言語の要求に基づき，データベース中で内容的に関連する文献の検索を行うシステムを研究中である。本システムにおいては，パーサが内容の比較の前段として重要な役割をもち，ユーザの質問文とデータベース中のテキスト文の両方の解析を担当している。本稿ではパーサの構成，処理について簡単に説明した後，大量のテキストの内容比較に関連した特質，問題点について述べる。

1．はじめに

我々は知的情報検索システムIRIS［杉山(1986)］の研究を進めている。本システムではユーザの自然言語による検索要求に基づき，内容的に対応するテキスト情報をデータベース中で検索することを目指している。システムの構成を図1に示す。

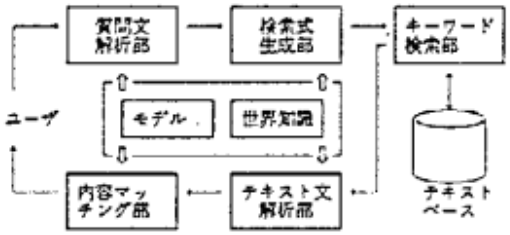


図1　IRISの構成

ここでは大量のテキストはあらかじめ解析され，内容表現とともにデータベースに登録されている。ユーザの質問文が入力されると内容表現に変換され，キーワード検索，内容照合の2段階の処理によって対応する文献の検索が行なわれる。

この処理の流れの中でIRISの自然言語解析部は内容の比較の準備段階として，質問文，テキスト文の内容表現への変換の処理を担当している。ところが，この2種類の文は同じ分野の内容を扱いながらも文体が大きく異なっている。また実時間でマンマシンインタフェースとしての処理をする質問文の解析部と大量のテキストをバッチ的に処理するテキスト文の解析部ではおのずから要求される仕様も異なってくる。このため我々は質問文解析部（Qパーサ），テキスト文解析部（Tパーサ）の2つの異なるパーサを作成した。両者は共通の枠組みに基づいているが，細部での処理内容は異なるものとなっている。本稿では情報処理産業界に関する新聞記事を対象としたプロトタイピングについて報告する。まず各パーザの構成について簡単に説明した後，質問文約100件，テキスト文約800件に対する評価結果について論じる。

2．パーサ概要

2．1．パーサの基本設計

パーサは基本的にbottom-up, deterministic な処理を行なう。これは品詞レベルでのあいまいさは少ないが文の構要素がはっきりとした順序をもたない日本語の特性，実行時間やメモリスペースに大きな制約をもつ自然言語インタフェースとしての使用を考慮して決定した。

システム全体はprologにオブジェクト指向の特徴を加えた言語ESP［Chikayama(1984)］によって記述されている。パーザの構成は基本的にはshift-reduce parserによっており，左側（あるいは右側）から順に単語間のまとめ上げの操作を繰返すことによって処理を進める。全体の処理は多段のフェーズに分割されており，各フェーズでの処理は一つのルール・オブジェクトが行なう。ルール・オブジェクトはESPのオブジェクトとして定義されており，ルール・インタプリタ部を継承している。またオブジェクト相互の継承を用いることによって共通する処理を行う部分の部品化を図っている。

2．2．対象分野のモデル化

IRISではシステムの枠組みを一般性のあるものとするため，対象分野に固有の知識を意味モデル，世界知識という形で独立して保持している。世界知識は，分野内で成立する事実に関する知識であるが，文解析処理では参照しないため，詳細については省略する。意味モデルは分野内の事物の構造をモデル化したものであり，主要な概念とその階層関係，さらに概念間に成立しうる意味的な関係がオブジェクト指向の枠組みで記述されている。

テキストベース中にはテキストそのもの以外にそれに付随する書誌情報（発行年月日，掲載された場所など）が記載されている。このため，意味モデルはテキスト内容を表す内容モデル，書誌情報を表す背景モデルのふたつから構成される多重モデルとなっている。IRISでは内容比較を一つ一つの事象を単位として行なうため，内容モデルは事象を特徴づけるような述語を中心とした構成となっている。背景モデルはそれに対して，データベースの編成項目である名詞的エントリを中心とした構成をとっている。

IRISのもつ辞書中の各単語（自立語）には各々の対応する可能性のある意味モデルのクラスへのポインタ（複数）が記述され，表層上の単語と意味モデルとの対応をとっている。

2．3．意味処理の枠組み

文の解析は表層の単語列を隣接要素間のまとめ上げ操作の繰り返しによって最終的に中心単語1つへと変換して行くことによって行われる。まとめ上げの操作の際には部分的な解析結果を意味モデルのインスタンスのネットワーク（意味木）として単語内部に保持する。その際，同じ構文構造をもつ（同じまとめ上げの操作を受けることに対応する）限り単語のもつ多義性は保存されるが，異なる構文構造を同時に扱うことはしない。意味的なチェックをする必要がある時は，システムがモデルに対してメッセージを送り，モデルがその結果を返すという形で行なう。

2．4．質問文解析部（Qパーザ）

QパーザはIRIS全体の中でユーザの質問を解析して，結果をテキストの検索部にわたす役目を負っている。対象とする質問文の例を下に示す。

例1　最近ＣＡＤ分野に進出した半導体企業について知りたい。

例2　先月のＡＴＴの提携に関する記事を示せ。

上に示すように質問文の多くは埋め込みによる修飾を多用した疑問文，命令文の形となっている。また，質問文の内容はテキスト内容に関する指定（内容モデルに対応），書誌情報に関する指定（背景モデルに対応），システムに対する要求表現（「～が欲しい」，「～を示せ」等の表現）を含んでおり，Qパーサはこれらの分離，抽出を同時に行う。Qパーサの構成を図2に示す。システムは文解析部，暗黙情報の解決部，意味モデル部の3つから構成されている。

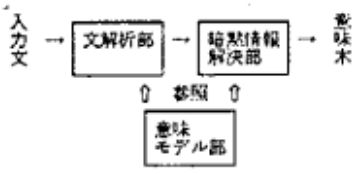


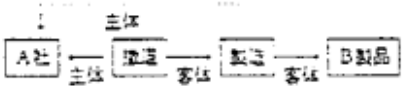
図2．　Qパーザの構成

内部での処理は2段階にわけて進められる。質問文はまず文解析部において各種情報の分離およびネットワーク構造への変換の処理を受ける。ここでの処理は前章で説明されたパーサの基本的枠組みを利用し，節の生成と節間の掛り受け関係の解析という順序で行われる。ここで節とは構文的曖昧さを持たない意味処理上の基本単位であり，「名詞連続＋助詞」，「述語＋（助動詞）＋（助詞）」等の構造のことをいう。

節の生成は意味モデルとは無関係に行われるが，節間の掛り受け関係の処理では意味モデルを利用して，意味的なチェックによる単語の多義のふるい分け，ネットワーク構造の生成を行う。またこの処理は2つのフェーズに分割されており，連体修飾節の合成，連用修飾節の合成の順で処理が行われる。連用修飾節の合成部においてはQパーザに特有な処理として，ユーザの要求を表す部分の分離を行う。

暗黙情報の解決部は前段での解析結果を受け，意味表現の標準化を行う部分である。意味表現同士の比較は基本的に述語とその必須格を対象として行われるため，比較の前に特に省略された格の特定をしておくこと（暗黙引数の解決）が必要となるわけである。質問文には多段埋め込み文が頻出し，それによる格の省略が多く見られる。

例　Ａ社はＢ製品の製造から撤退したのか？



例えば上の例では，「製造」の主格が明示されていないが，主文の主格である「Ａ社」がこの述語の主格を占めることは明白である。Qパーサの暗黙情報の解決部では上のように同じネットワークの中で対応する部分を探すことによって省略された格の解決を図っている。

2．5．テキスト文解析部（Tパーザ）

Tパーザはテキストのデータベースへの登録の際に内容表現を抽出するためのシステムである。今回のプロトタイピングではテキストとして処理コストの高い新聞記事の本文ではなく，記事見出しを扱っている。

新聞の記事見出しの例を下に示す。（◆は見出し間の区切を示す記号）

例3　理経◆ＤＥＣ互換機を拡販◆米ク社と代理店契約

例4　日本ＣＤＣ◆非線形解析ソフト発表◆処理時間１／１００に短縮

例5　韓国ソウルエレクトロン◆ＣＡＤＣＡＭで米ＣＶと合弁会社

これらの例からわかるように，記事見出しは，

ⅰ）　助詞，動詞語尾などの省略が多い。特定できる場合は動詞本体まで省略される。

ⅱ）　一般に主見出し，副見出しといった複数の区画から構成されている。

ⅲ）　基本的に複数の文から構成される文章であり，文同士で頻繁に格の共有が行われる。

等の特徴をもっている。また実際の紙面上での理解を助けている平面的なレイアウトに関する情報はデータベース登録時には，一次元の構造に還元されてしまうため，さらにあいまいさを増やす結果になっている。

これに対し，Tパーサでは，基本的な構成，アルゴリズムは，Qパーサと同一のものを使い，付加的な意味チェックの追加，暗黙情報の解決部で文間の関係を解析すること等によって対処している。またここで対象とする文はテキスト本体だけなのでモデルとしては内容モデルのみを参照して解析を行う。

文解析部では一つの区画内を対象としてまとめ上げの処理を行う。基本的な処理の組立はQパーザの文解析部

と同様であるので繰り返さない。ここでは変更部分についてのみ述べる。

［名詞連続部の処理］

助詞等の省略は互いに無関係な語の並びを名詞連続にみせるため，名詞連続の範囲認定がむつかしくなる。このため名詞連続の処理の際には意味的情報を導入して単語同士が名詞連続として連接し得るかのチェックを行なっている。

［サ変名詞の扱い］

サ変動詞の語尾省略によって発生する品詞のあいまいさに対応するためにサ変名詞を述語としても扱っている。

暗黙情報解決部では，区画間の関係の解析を担当している。ところが区画の統語的な役割は一定していない。区画の構成はレイアウト上の都合によって決定され，各区画が１つの文に対応する場合や文内の１つの節に対応する場合など様々な場合がある。これは区画間に文内のかかり受け構造や，複数の文間の格の共有などの多様な関係が存在することを意味する。このため暗黙情報解決部では図３に示すような多段にわけた処理を行う。

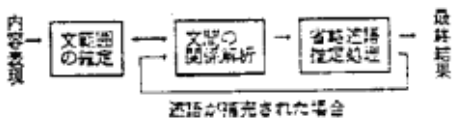


図３　暗黙情報の解決部での処理の流れ

まず文範囲の確定部では文内でのかかり受け関係の処理によってその後の処理の基本単位である文のまとめ上げを行う。文間の関係解析部では各文内の述語の未解決の必須格に対し，候補を記事全体の走査によって見つけ出す。最終段の省略述語の推定処理部では，これまでの処理でかかり受け先の決定できなかった連用修飾節について省略された述語を推定する。例えば「Ａ氏がＢ社社長に」といった文において「就任する」という本動詞の推定を行う。

３．評価

IRISの場合，解析の最終目標は（書誌情報の部分を除けば）抽象的なテキスト内容に関する指定であるため，受理すべき文の範囲や解析結果の正当性の判断が難しくなり，解析の成功／不成功がはっきりと決定できない。このため解析結果の評価を次のようにある程度連続的なものとした。

(1)　成功　……　目標とする構造が生成される

(2)　不完全……　誤りではないが，目標とする構造には情報が不足している（リンクの欠如等による）

(3)　失敗　……　構造の生成に失敗する，あるいは目標とする構造と明らかに違う構造が生成される

３．１．質問文解析部の評価

質問文解析部についてはアンケート調査から典型的なものとして抽出した100 例の質問文を対象とした実験を行った。抽出の際の基準としては，

(1)　主な自立語がモデルの範囲にはいっていること。

(2)　記事内容で述べられている事実に対する質問であること。

の２つだけを考え，特に文型などに制限を加えることはしなかった。

これらの文例に対し，辞書エントリがすべて整った状態での最終的な解析結果は，

(1)成功…79%　(2)不完全…　9%　(3)失敗…12%

となった。

不完全／失敗の原因としては，(1)ルール適用アルゴリズムの不備，(2)格関係に関する意味条件，統語条件の不適，(3)モデルの表現能力の不足，(4)代名詞の照応関係が未解決，(5)特定の文型の質問文に対応するルールの不備，などが挙げられる。

インタフェースとしてのパーサの機能は実際にはシステムの他の部分の働きと密接に関連している。基本的にはテキスト情報の検索しかできないシステムにたいして複雑な判断を求めた質問文を入力しても適切な対応は期待できない。その意味ではQパーサの場合，受理すべき文の範囲及びその範囲内での解析結果をはっきりと規定することが必要である。

我々はこの意味から複文，重文における文関係の解析に踏み込むことはやめている。そして時間表現の様々なバリエーションに対応するために背景モデルの拡張を行なっている。

今後はシステム全体としての機能向上と合せて，文脈処理の導入，様々な照応表現への対応を図って行く考えである。

３．２．テキスト文解析部の評価

産業新聞３紙（日経産業新聞，週間電波コンピュータ，日本情報産業新聞）から情報産業界に関連した記事を約800文抽出したものをターゲットとして辞書および文法規則類の整備を行なった。

現在のデータベース中の文献に付加されている内容表現についての評価値を下に示す。（ただしこれらの内容表現はパーサの改良を行いながら順次登録して行ったものであり，古いバージョンのパーサでの解析結果が含まれている。このため現在のパーサの能力の公正な評価とはならないことをあらかじめことわっておく。）

(1)成功…60%，(2)不完全…15%，(3)失敗…25%

不完全，及び失敗の主な原因としては，(1)名詞句の範囲認定の失敗，(2)格関係の処理の失敗，(3)照応関係の解決の失敗，などが挙げられる。

名詞句の範囲認定の失敗は，名詞連続部の認定での処

理の失敗，等位接続の範囲認定の失敗等が多い。格関係の解析誤りは，語義の多義性，クラス分けの詳細度の不足などが原因となっている。

照応関係の解決処理では最初，処理間違いが頻出したため，処理を必要最低限のものに限っている。このため現状での失敗原因の多くを占めてはいないが，パーサでの処理の精密化に対する大きな障害となっている点に変わりはない。処理における問題点としては，

i） 照応関係は述語間の格の直接の共有以外にも様々なタイプのもの（例えば製品名とその属性の間の照応等）があること。

ii） 述語間の格の共有関係でもテキストの構造，実世界の関する知識なしでは正しい候補の選択ができないものが多い。

等が挙げられる。

ここに現れる問題へのTパーサ側での対応としてはまず照応処理の改良を考えて行きたい。テキストの構造を認識してそれに応じた柔軟な探索が行なえるようにすることが必要だが，それに併せた分野知識の整備の支援ツール開発も緊急の課題である。

4．考察

プロトタイピングによって明らかになった一般的な問題点についての考察を述べる。

4．1．分野知識の不足

もともと内容表現同士の類似度の比較という目的からみれば，本システムには分野知識が不可欠のものではあるが，文解析の処理においても，統語上の手がかりの少ない部分の処理では分野知識への依存度が大きくなっている。ところが述語を中心として考えられてきた従来の内容モデルでは，名詞的概念の構造に関する知識が不足している。特に見出し中に頻出する製品に関する知識の不足は，照応関係の解析や，等位接続の処理上の問題となる。このような知識を早急に整備することが今後の重要な課題である。

ただ大規模のモデル（現状でのクラス総数は約160 ）の作成と保守はそれ自体，非常に難しい問題である。特に内容モデルの場合，モデル内のエントリに直接に対応するものがなく，クラス分けの妥当性は内容比較の処理がうまく行くかといった間接的な要因によって判断せざるを得ない。

結局大量のテキスト情報の処理をする限り，分野知識が常に完全であることは期待できないだろう。扱う分野が多少異なる文章がはいってくるかもしれないし，同一分野内でも技術的な内容は年々変化して行く。文解析部としては不完全な知識の影響をできる限り，局所化するようなものとすることが重要である。現在のところは格関係の判定条件に関する不備が多いため，格関係の処理の際にかかり受けの距離と意味的な妥当性を合せて判断することによってある程度対処している。

4．2．環境整備

一般の自然言語インタフェースと違い，本システムでは大量のテキスト文に対し，意味的な解析を行わなければならない。またテキスト文の解析に際しては受容する文の範囲をあらかじめ規定するわけにはいかない。このため辞書整備，文法の検証にかかる手間は膨大なものとなり，システムを運用してゆく上での大きな問題となってくる。実用的なシステムを構築して行こうとすればこれらの作業の支援するための方策が必要となってくるだろう。

辞書整備に関しては，特に変動の激しい社名，製品名などの固有名詞の整備が一番の問題となる。これに対しては次のような方策を考慮している。

i） 入力文の単語分割の際に，製品，あるいは会社名の一般的なパタンに合致するものを探すことによって辞書中にない固有名詞を認識する。

ii） 記事本文には新出の語句に対する説明が含まれることが多い。このため，本文の解析によって固有名詞に関連する情報の抽出を行う。

一方，文法の検証に対する環境整備に関しては，

i） 複雑なネットワークの表示を最適化するためのツール類を整備する。

ii） 解析時にネットワークに適用ルールの情報等を付加し，後にオフラインである程度問題箇所の特定ができるようにする。

iii） 文法規則のモジュール化を進めることによってパーサ間での部品の共有を図る。

等を検討中である。

5．終りに

現在IRISではテキストベースの大規模化と分野移行性の検証［秋山1987］へ向けての作業を行なっている。システム全体としては分野知識の整備への支援が一番の課題となっている。文解析部としては，大量のテキストの解析に耐えうるように文法規則を強化するとともに，システムの構築，保守のためのツール類の整備にも力を注いで行くつもりである。

謝辞 本研究は，第５世代コンピュータ・プロジェクトの一環として行われた。本研究に対して御支援頂くICOTの方々に深く感謝致します。

参考文献

［杉山(1986)］杉山他 "自然言語理解に基づく情報検索システムIRIS" 情報処理学会ＮＬ研資料58-8.1986

［Chikayama 84］Chikayama, T. "ESP Reference Manual", ICOT Technical Report TR-044, 1984.

［秋山(1987)］秋山他 "知的情報検索システムIRISの分野移行性の評価" 情報処理学会第35回全国大会予稿集 pp1429-1430, 1987.