自然言語インタフェースとしてのIRIS

伊吹　潤，杉山健司，鈴木香緒里，玉田億子，川崎正博
（富士通）

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。これは自然言語テキストの大規模データベースへのアクセスを容易にするための知的インタフェースをめざすシステムである。ここではユーザの自然言語による質問を受けつけ，質問内容に意味的に合致するような内容をもつテキストの検索を行なう。

この処理の中でIRISの自然言語インタフェース部は，ユーザの質問文の解析，及びテキスト文の解析の2つの働きを行なう。いわばテキスト内容表現という同一の目標に対してユーザの検索質問文，テキスト文という2つの異なるタイプの文からの変換を行なうわけである。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り，我々はそれらの解析のために質問文解析部（Qパーザ），テキスト文解析部（Tパーザ）の2つの異なるパーザを作成した。本稿では新聞記事を対象としたプロトタイピングの事例について報告する。まず各パーザの構成について簡単に説明した後，サンプル文に対しての実験結果について論じる。

2．パーザ概要

2．1．パーザの枠組み

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関係によって共通する処理を行なう部分の部品化を図っている。

2．2．意味処理の枠組み

本システムでは分野に依存する知識を意味モデルという形で独立して保持している。これはオブジェクト指向の枠組みで記述され，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定義されている。

IRISではシステムに対する質問としてテキスト内容以外にも発行日，掲載新聞名などの書誌情報に関する指定もとり扱っており，意味モデルではそれぞれテキスト内容を表す部分を内容モデル，書誌情報を表す部分を背景モデルとして別のモデルとして扱う。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）が記述されている。

文の解析は表層の単語列を解析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。

2．3．質問文解析部（Qパーザ）

Qパーザはユーザのシステムへの質問を解析して，テキスト内容に関する指定（内容モデル）と書誌情報に関する指定（背景モデル）を分離，抽出する。構成は文解析部，意味モデル部，暗黙情報の解決部の3つからなっている。

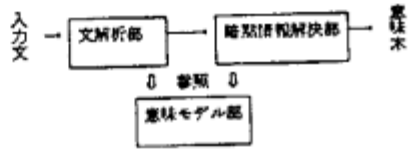

図1．　Qパーザの構成

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ザの基本的枠組みによって行われるが，ここでの操作の基本は節の生成と節間の係り受け関係の解析である。ここで節とは構文的，意味的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，節間の係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，「〜を示せ」，「〜を教えて下さい」といったユーザの要求を表す部分の分離を行う。

暗黙情報の解決部では，述語の暗黙引数（他の情報から特定できるために省略されている格）に対し，意味木の構造の中に対応する部分を探すことによってその解決を行っている。Qパーザの場合は述語の多段埋め込みによる主格の省略を扱う。例えば「A社はB製品の製造を開始したのか？」という文において，構文上は現れない「製造」の主格を意味木の中の探索によって「A社」と特定するような処理を行う。

Natural Language Interface Part of IRIS
by Jun IBUKI, Kenji SUGIYAMA, Kaori SUZUKI,
Ikuko TAMADA and Masahiro KAWASAKI
Fujitsu Ltd.

2．4．テキスト文解析部（Tパーザ）

Tパーザは新聞記事のデータベースへの登録の際，記事見出しの解析によってテキスト意味表現を自動抽出するためのシステムである。

Tパーザの全体の構成はほぼQパーザの場合と同じであるがここで対象とする文はテキスト本体のみ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。

文解析部での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が，省略表現が多く表層上の手がかりの少ない見出し文の解析のために各部での処理の細部がQパーザと異なる。特に助詞等の省略は互いに無関係な節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。また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。

見出し文は一般に主見出し，副見出しといった複数の区画から構成されている。暗黙情報解決部では，こういった区画間の関係の解析を行なう。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まず文内のかかり受け関係の処理によって文のまとめ上げを行い，その後に文間の格の共有関係の処理を行なうこととしている。またこの処理によって掛り受け先の決定できなかった連用修飾節については本動詞が省略されたものとして，省略動詞の推定の処理を合せて行う。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むつかしくなり，解析の成功／不成功がはっきりと決定できない。例えば極端な場合，すべてのかかり受けの解析に失敗してフラットな構造が生成されたとしよう。この場合，システムはキーワード・レベルの検索を行うこととなってしまうが，この場合でもある程度の結果が保障されるわけである。したがって解析結果の評価について次のような判断規準を定める。

(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

まず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。扱うべき文の範囲としては，

(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つを規準とした。

上で述べた規準をあてはめると，100文中(1)成功…79例(2)不完全…9例(3)失敗…12例，となっている。

不完全／失敗となる理由は，(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。

パーザ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。パーザの機能のさらなる向上のためにはパーザの枠組みに対する検討のみならず，モデルの表現形式の改良，談話状況を認識した文脈処理の導入など基本的なシステムの枠組みに対する検討が必要となる。

3．2．テキスト文解析部の評価

対象として産業新聞3紙から情報産業界に関連した記事を72文抽出したものを用いて実験を行った。

現在の状態では72文中(1)成功…50例，(2)不完全…11例，(3)失敗…11例となっている。不完全，及び失敗の主な原因としては，(1)モデルの情報不足（モデルの詳細化が必要），(2)モデルの表現能力の不足，(3)未登録語処理アルゴリズムの不備，などが挙げられる。

スペース上の制約の多い見出し文には複雑な統語構造が現れることはまれである。このため文解析のアルゴリズム上の問題は割と少ない。しかしながら統語上の特徴の少ない見出し文の解析のためにはテキストに述べられている内容についての様々な知識が必要であり，これらをどう整備すべきかが大きな問題となる。

4．おわりに

テキスト・データベースを扱う自然言語インタフェースとしての問題点を各パーザの場合について述べてきた。これらは技術上の問題であるが，実際にシステムを運用していく上では単語辞書のエントリが数千の規模となり，整備に多大の手間がかかる点が大きな問題となる。このため，我々は変動の激しい製品，企業に関する情報を記事本文の解析から自動抽出する可能性についての検討を開始している。

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の横井俊夫，岩下安男両室長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"　情報処理学会自然言語処理研究会資料58-8. 1986

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