知的情報検索システムＩＲＩＳにおける等位接続解析方法について

鈴木　香緒里，秋山　幸司，川崎　正博
富士通株式会社

1．はじめに

等位接続句のあいまい性をいかに解決するかが，さまざまな自然言語処理システムで問題となっている．日本語質問文を理解し，その解答となるテキスト群をテキストベースより検索する知的情報検索システムＩＲＩＳ（Intelligent information Retrieval and Information Sellective System)の自然言語インタフェイスである質問文解析部でも，ユーザの質問を正しく理解するうえで，等位接続のあいまい性が問題となっている．等位接続は等位な関係にある要素が格助詞「と」やそれに相当する語で結ばれている句で，それらの要素は，格助詞「の」またはそれに相当する語で修飾されている場合もある．

本稿では，質問文解析部で採用した等位接続句解析の一手法について述べる．

2．質問文解析部の概要

質問文解析部では，図１に示すように形態素解析の後，構文，意味情報を用いてまとめあげ処理を行い，構文意味木を生成する．意味モデルは，対象世界に属する概念クラスを記述する．述語概念と名詞概念との格関係，名詞概念間の関係，等位な関係になりうる概念の組み合わせなどの情報を持ち，解析の際に参照される．等位接続の解析では，構文情報が少ないため，意味情報が重要となる．このため，等位接続が正しく解析されるか否かは意味モデルの記述の正しさにかかっている．

また，助詞を介さずに名詞が続く名詞連続と格助詞「の」などを介した場合とを同等に扱っている．従って，「生産計画」は一語として扱わずに，「生産の計画」と同じく，「生産」が「計画」を修飾しているとして解析を行う．

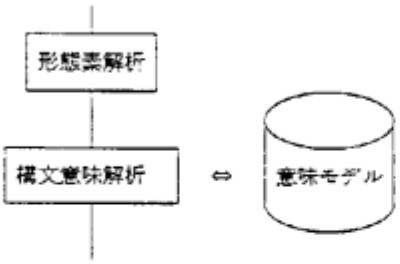

図１．質問文解析部の概要

3．等位接続の解析方法

これまで，質問文解析部では，入力・出力バッファと係り先がまだ解決されていない要素を保存するスタック（以後，作業用スタックと呼ぶ．）を用いて解析を行っていたが，等位接続の解析時に以下のような問題があった．

たとえば，「Ａ社のＭＳＸパソコンと１６ビットパソコン」という名詞句を解析する場合，解析は左から順に決定的に進めていくので，等位な関係にある前の「パソコン」と後の「パソコン」をまとめようとする時には，すでに「Ａ社の」は前の「パソコン」とまとめられ，文の表層からは消えてしまっている．このため，「Ａ社の（（ＭＳＸパソコン）と（１６ビットパソコン））」という意味なのに，「（Ａ社のＭＳＸパソコン）と（１６ビットパソコン）」という意味の構造ができてしまう．

このような問題を解決するために，新たに二つのスタックを導入した．一つは，等位な関係にある要素を保存するための等位要素用スタック，もう一つは，それぞれの等位な要素に係る修飾句を保存するための修飾句用スタックである．これは，修飾句とそれが修飾する等位要素との対応を記憶しておく必要があるので，一つの等位要素の修飾句に対して一つのスタックが割当てられており，これらのスタックが等位要素用スタックの要素である．つまり，修飾句用スタックはスタックの二重構造になっている．

次に，例の等位接続句を用いて解析手順を説明する．

例）「Ａ社のＭＳＸパソコンと１６ビットパソコン」

解析は左から順に行う．解析開始時の状態を図２ａに示す．

① 入力バッファの先頭には「Ａ社の」という要素があるので「Ａ社の」と「ＭＳＸ」とのまとめあげを試みるが，意味モデルにこれらの間の関係の記述がないので失敗し，「Ａ社の」は作業用スタックに入る．

② 「ＭＳＸ」と前の等位要素である「パソコンと」が入力バッファの先頭に現れるが，「ＭＳＸ」は後ろの等位要素にもかかる可能性があるので，この時点ではまとめあげ処理は行わずに「ＭＳＸ」を作業用スタックに，「パソコンと」を等位要素用スタックに入れる．

③ 等位要素スタックに入っている要素と等位な関係にある要素を探す．まず，「１６ビット」と「パソコンと」が等位かを調べるために意味モデルを参照する．その結果，等位な関係ではないことがわかるので，「１６ビット」は後の等位要素の修飾句と判断できる．

④ 同様に「パソコン」について調べると，これが後の等位要素であることがわかるので，これを等位要素スタックに入れ，「１６ビット」は修飾句用スタックに入れる．この時の状態が図２ｂである．

⑤ 処理④で等位要素の探索に成功したので，次に，修飾句用スタックの内容と等位要素との連体修飾のまとめあげ処理を行う．この際，まとめあげようとする要素をどのような関係で結ぶかは，意味モデルに記述されている，対応する概念クラス間の関係定義による．

⑥ 作業用スタックの内容と等位要素との修飾関係を調べる．作業用スタックには，前の等位要素のみに係るものと等位要素全体に係るものの両方が存在可能である．この例では「Ａ社の」が全体に係り，「ＭＳＸ」が前の等位要素のみに係る．これを調べるには，意味

The Method of Japanese Parataxis Analysis in IRIS

Kaori SUZUKI, Kohji AKIYAMA, Masahiro KAWASAKI

Fujitsu Limited

図２ａ．等位接続解析開始時の状態

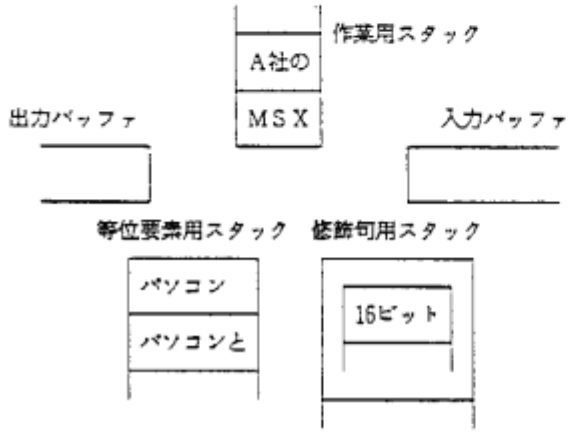

図２ｂ．解析途中の状態

図２ｃ．解析終了時の状態

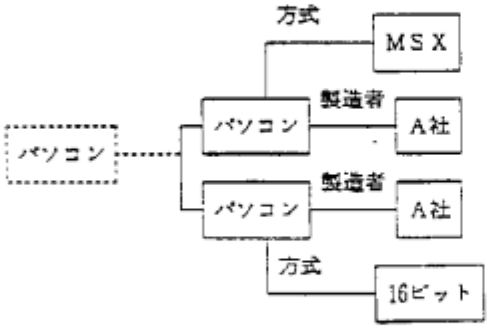

図２ｄ．生成された構文意味木

モデルを参照し，これらと等位要素との関係をチェックする．「ＭＳＸ」は前の等位要素「パソコンと」の形式を表すことがわかるが，後の等位要素「パソコン」はすでに「１６ビット」という形式を表す要素で修飾されているので，「ＭＳＸ」は前の等位要素のみを修飾することがわかる．一方「Ａ社の」は両方の等位要素と製造者を表す関係にあるので，等位要素全体に係ることがわかる．等位接続解析終了時の状態を図２ｃに，生成された構文意味木を図２ｄに示す．図２ｃの出力バッファに現れた「パソコン」は，解析の都合上生成したダミーのインスタンスであり，図２ｄの点線に囲まれた要素に対応する．図２ｄ中の四角はインスタンスを，直線はインスタンスを結ぶアークを表し，添字はアーク名である．

4．実験結果

情報産業関連の新聞記事本文より等位接続句の例 117 件を抽出し，本手法による解析実験を行った．その結果次のような問題が現れた．

①修飾の範囲の誤り．

例）「大企業と計算センター」を「大（企業と計算センター）」と解釈する．これは，格助詞「の」を介した場合と介さない場合を同等に扱っているためであり，格助詞の有無によって修飾しうる範囲を変化させる必要がある．

②等位な要素の選択の誤り．

例）「交換機とデータ端末機」を「（交換機とデータ）端末機」と解釈する．これは，「交換機」と「端末機」は”製品”という概念クラスに，「データ」は”データ”という概念クラスに属しており，意味モデルには”製品”と”製品”，”製品”と”データ”の両方の組み合わせが等位接続可能と記述されているので，先に現れる「データ」のほうが後の等位要素と判断されてしまうからである．これに対しては，名詞の連語同士が等位接続しやすいといったパターンを利用したり，等位接続する概念クラスの組み合わせに，その結びつきの強さによって優先順位を付けることなどで対処したい．

5．おわりに

本手法により，これまでの質問文解析部では構造上，解析不可能だった，等位な要素全体にかかる修飾句を含む等位接続句が解析可能となった．このような等位接続句の割合は，今回抽出した例中では23%を占めていた．

今後は，①，②で示した問題を解決することで等位接続句解析能力の向上を図る予定である．

謝辞　本研究は第五世代コンピュータプロジェクトの一環として行われた．御支援いただいたICOT関係者各位に深く感謝いたします．

参考文献

〔1〕杉山他：”自然言語に基づく情報検索システムＩＲＩＳ”，情処NL研58—8，pp. 1-8，1986

〔2〕伊吹他：”自然言語インターフェイスとしてのＩＲＩＳ”，情処34回全国大会4X-8，pp. 1325-1326，1987