オブジェクト指向パーザ POWER による日本語構文解析とその評価

高橋博之　　宮崎正弘

新潟大学大学院自然科学研究科

1 はじめに

オブジェクト指向パーザ POWER[5] は、手続き中心の記述により柔軟な記述が行なえるパーザである。

我々はこの POWER を用いて日本語文パーザの実装を行なった。まず従来の研究を元に表層的な解析を行なうパーザを作り、それに POWER の継承の機能を利用して、個別的な意味解析による細かな制約を実装した。

最後に評価を行なった。評価には新聞の社説などを用い、係り単位で約 97 %、文単位でも約 60 % の解析率を達成した。

2 オブジェクト指向パーザ POWER

オブジェクト指向パーザ POWER は、オブジェクト指向の考え方を元にした、柔軟で記述性のよいパーザである。POWER では単語に対応する実体である単語オブジェクトが互いにメッセージ通信で情報を交換して解析を進める。

各単語オブジェクトはクラス変数と呼ばれるローカルデータを持つが、他のオブジェクトはそれを直接読み出すことができず、メッセージ通信によって問い合わせる必要がある。このことによりデータのカプセル化が行なわれ、データの依存関係がはっきりする。

POWER では別の単語オブジェクトからの継承で新しい単語オブジェクトを作ることができる。この場合、元の単語の機能はそのまま受け継がれるので、それに新しい動作を加えるか、あるいは従来の動作を上書きすることで、少ない記述で新しいオブジェクトを作成することができる。例えばオブジェクト「動詞」に細かい格制約を追加してオブジェクト「走る」を導出できる。

3 従来の係り受け解析との違い

3.1 文節

従来の構文解析では「文節」を解析の単位として、文節間の係り受けを解析していた。しかしこれだと意味的に不自然な場合が出てくる [7]。

例えば「花が咲くか」と「花が咲くだろう」は文節で切るとそれぞれ「(花が) (咲くか)」、「(花が) (咲くだろう)」となる。しかし、これを意味的なまとまりから考えると「(花が咲く) か」、「(花が咲く) だろう」と区切る方が自然である。このため本パーザでは文節という単位は使用せずに、付属語も独立した単語オブジェクトとして実装している。

3.2 係り受けの交差

従来の係り受け解析では係り受けの交差を認めなかった。この制約は多くの場合妥当である。しかし、例えば「うなぎを浜松に食べに行く。」という文では明らかに交差した係り受け構造が自然である。

Japanese syntactic analysis with POWER, and evaluation.

Hiroyuki Takahashi (hiro@tinlp.info.eng.niigata-u.ac.jp),
Masahiro Miyazaki (miyazaki@info.eng.niigata-u.ac.jp)
Niigata University

一方、実験によって係り受けの交差を明示的に禁止しなくても後述する係り先の制約によって、係り受けの交差が誤って生成されることはほとんどないことがわかった。そのため、POWERでは係り受けの交差を認めることにした。

4 解析の概要

本パーザは接続優先度に基づいた単純な係り受け解析を元に、個別的な処理を細かく記述していくという形で作られている。

4.1 係り先の探索

本パーザでは各単語がメッセージ通信によって係り先を探す。単語が探索のために送るメッセージは様々で、そのメッセージの種類によって探索対象が決まる。例えば副詞は「連用修飾」メッセージを送信し、それを用言が受け取って係り関係「連用修飾」を生成する。この時、間にある名詞などの連用修飾と関係のない単語はこのメッセージを通過させ、探索には関与しない。メッセージは語順にそって流れるので、基本的には一番近い用言に係る。

各種探索メッセージによる探索元、探索対象対は [1][3] [2] における係り規則とほぼ同様のものである。ただし、例えば形容詞は「を」格をとらないなどの細かい制約は受け側のメソッドに記述されている。形容詞は格助詞からの探索メッセージを受け取るとその格のチェックを行ない「を」格の場合はメッセージを通過させる[1]。するとこのメッセージは次の用言へ流れ、その用言がまた制約をチェックするというように進行し、最初に制約を満たす用言に係る。

4.2 接続優先度

助詞「は」は助詞「が」などに比べ、比較的長距離を係る傾向がある。また一般に読点を伴う句や節も比較的長距離の係りになる。このように表層的に決まる係りの長さを整理したものが接続優先度である。これは [6] における従属節間の係りの制約をもとに、これを一般化して格後置詞や副詞にも対応させたものである。

接続優先度を表 1に示す。接続優先度は係り関係に以下の制約を与える。

[1]正確には一度受け取ったメッセージを再送信する

（係り側の優先度）≤（受け側の優先度）

つまり接続優先度が高いほど係ることができる単語が少なくなり、より長く係るようになる。

優先度	係り側	受け側
6	「展開」の節＋読点	主節 「展開」の節＋読点
5	「展開」の節 「引用」の節 助詞「は」＋読点	「展開」の節 「引用」の節
4	「条件」の節＋読点 連用中止節＋読点 助詞「は」 格後置詞＋読点 副詞句＋読点	「条件」の節＋読点 連用中止節＋読点 形式名詞に係る連体節
3	「条件」の節 連用中止節 副詞句	「条件」の節 連用中止節 普通名詞に係る連体節
2	「同時」の節＋読点	「同時」の節＋読点
1	「同時」の節 格後置詞	「同時」の節

表 1: 接続優先度

5 個別的処理

前述のような一般的な規則に基づく単純な係り受け解析ではどうしても対処できない場合がある。このような例外にはクラスの導出を利用して対処することができる。

5.1 助詞相当語

日本語には「について」など、動詞の連用形を使った助詞相当の表現がある。多くの場合は、このままの形で助詞相当語として辞書に入れるか、あるいは形態素解析結果に後処理をかけてまとめれば済む。しかし、例えば「をめぐる」の場合、「利権をめぐって争う」という助詞相当の用法のほかに「美術館をめぐった」というような普通の動詞としての用法もある。この判別は名詞のカテゴリによってある程度行なうことができる。そこで本パーザではこの動詞「めぐる」のクラスを特別に作り、この判定を行なう。このクラスは「動詞」クラスからの継承で作られている。そのため、一般の用法と判定された時には、「動詞」クラスの記述をコピーしなくても普通の動詞としてのふるまいを行なうことができる。

5.2 形態素の曖昧性解消

形態素解析では表層的な接続のみで解析するので、どうしても解消できない曖昧さがある。一般には適当な経験則で一つに絞りこんでから構文解析を行なうが、もし形態素解析ですでに誤っていたら構文解析は全く無駄になってしまう。本パーザでは構文解析が誤りがちなところに曖昧性をもったオブジェクトを配置することでこの問題に対処する。

例えば助詞「まで」は格助詞と副助詞（強調）の用法があり、形態素解析では通常その判別ができない。そこで「まで」に対しては格助詞と副助詞の両方の機能を持ったオブジェクトを生成し、そのオブジェクト内部に実装された判定処理でどちらの機能を作用させるかを決定する。この判別には主に名詞のカテゴリ情報を使い、「時間」や「場所」などを受ける場合は格助詞とし、それ以外は副助詞というように判定する。

5.3 名詞の並列構造の抽出

名詞の並列抽出はしばしば専用の処理段階を設けて行なわれる [1] [2]。しかしこのような方法では名詞の並列解析に構文解析の情報が利用できない（あるいはこの逆）という問題点がある。この問題は構文解析途中で随時並列抽出を行なうことで解決できる。

本パーザは以下の3つのステップで解析中に名詞の並列構造抽出を行なう。

1. 形式による抽出
2. 意味的類似による抽出
3. 動詞による後処理

まず形式による抽出では形式的に明らかな並列を抽出する。例えば「彼と私とで」というような「～と～と」パターンや、読点で区切って名詞を並べたものがこれにあたる。

次に意味的類似による抽出を行なう。これは名詞の後に助詞「と」「か」「や」が来た時に行なわれる。これらの助詞は後方の名詞でカテゴリの類似度が一定以上のものを探し、その名詞と並列にする。

次に動詞が処理を行なう。助詞「と」は必ずしも並列を示すわけでなく、例えば「彼を先生と呼ぶ」というように格を表すこともある。そこで、「と」は動詞を飛び越えて並列先の検索はできないこととし、動詞は自分が「と格」をとるかをチェックし、とらなければ適当な格要素と並列関係を作る。これは前の2段階で並列構造を抽出できなかったときのための対策である。

6 評価

新聞の社説などから収集した211文を使用して評価を行なった。使用した文の詳細を表2に示す。これらはあらかじめ形態素解析によって単語分割が行なわれている。係り単位での評価結果を表3に示す。これを見ると文の種類によらずほぼ一定の正解率を達成していることがわかる。

また、文単位の評価結果を表4に示す。文中の全ての係りが正しく求められた文のみを正解としている。文単位での正解率は60 %弱にとどまっているが、表5に示すように誤りの文のうち8割以上が一つか二つ程度の誤りであり、ほとんどの文でおおよその構造はつかめている。

出典	文の数	平均文字数
朝日新聞社説 ('97 12/11) 全文	69	38
日経新聞社説 A('97 8/26) 全文	35	48
日経新聞社説 B('97 9/7) 全文	57	40
「電子図書館」[4] 第1章前半	50	64
計	211	47

表 2: 実験に使用した文

出典	出現数	正解	正解率
朝日新聞社説	1114	1079	96.9 %
日経新聞社説 A	739	718	97.2 %
日経新聞社説 B	1021	986	96.6 %
「電子図書館」	1297	1247	96.1 %
計	4171	4030	96.6 %

表 3: 評価結果（係り単位）

6.1 解析の誤り

名詞句に関連するもの 名詞句に関連する誤りが一番多く、全体の半分近くを占めた。まず「魅力のあるところ」というような「名詞1 の 用言 名

出典	文の数	正解	正解率
朝日新聞社説	69	46	66.7 %
日経新聞社説 A	35	22	62.9 %
日経新聞社説 B	57	32	56.1 %
「電子図書館」	50	24	48.0 %
合計	211	124	58.8 %

表 4: 評価結果（文単位）

誤りの数	0	1	2	3	4	5
該当する文の数	124	53	21	7	5	1

表 5: 誤りの数の分布

詞 2」というパターンがある。現在のところ本パーザでは常に「名詞 1」から「名詞 2」への係りを優先するので、この例のような場合には対処できていない。また同様の誤りに「異なった種類のサービス」というような「用言 名詞 1 の 名詞 2」というパターンがある。これも「用言」から「名詞 2」への係りを優先するのでうまくいかない。

これらの問題は名詞句に関しての細かい意味的制約がまだ実装されていないことに原因がある。名詞句に関する誤りは局所的な問題であり、問題の設定が明確（A に係るか B に係るかの二者択一）であるので、今後、係り先を選択する精密な意味的制約規則を実装することによって対処できると思われる。このような場合、POWER では手続きを使えるので記述しやすい。

大局的構造　本パーザは各オブジェクトが局所的に動作するので大局的な構造を把握しにくい。例えば「これが電子メールと呼ばれているものだ」のような場合、「これが」に長距離係る要因がないので「呼ばれる」に係ってしまう[2]。このような大局的な構造を扱うためには、単語に対するオブジェクトに加えて構造に対するオブジェクトを作り、そのオブジェクトが大局的構造を抽出し各単語オブジェクトに通知するような仕組みが必要になるだろう。

[2]「これ」＝「もの」で、「もの」が「呼ばれる」のだから意味的に必ずしも間違いとは言えない

7　おわりに

オブジェクト指向パーザ POWER を用いて日本語文パーザの実装を行ない、評価実験により高精度の解析ができることを示した。

今後の課題としては、POWER は局所的な処理は簡単に記述できるが大局的構造の抽出のような処理を記述しにくいので、この点の改良が挙げられる。また名詞句の意味的制約が不十分であり、この点の強化も必要である。また、今後さらにさまざまな文章のより高精度の解析を目指し、個別的な記述を増やしていく予定である。

謝辞

本パーザの実装には NTT の意味属性体系データ、結合価パターン辞書を使用した。これらのデータを提供して下さった NTT コミュニケーション科学研究所の機械翻訳研究グループの方々に感謝する。

参考文献

[1] 稲垣, 壁谷, and 小橋. 意味連結パターンを用いた係り受け解析. In 情報処理学会自然言語処理研究会 *88-NL-67*, 1988.

[2] 黒橋禎夫 and 長尾眞. 並列構造の検出に基づく長い日本語文の構文解析. 言語処理学会論文誌, 1(1):35–57, 1994.

[3] 山上晃司 and 安原宏. 形態素情報による日本語の係り受け解析. In 情報処理学会自然言語処理研究会 *93-NL-98*, pages 9–17, 1993.

[4] 長尾眞. 電子図書館. 岩波書店, 1994.

[5] 高橋 博之 and 宮崎 正弘. オブジェクト指向パーザ power. In 自然言語処理学会第 2 回年次大会発表論文集, pages 157–160, 1996.

[6] 白井, 池原, 横尾, and 木村. 階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度. 情報処理学会論文誌, 36(10):2553–2361, 1995.

[7] 北原保雄. 日本語の世界 6 日本語の文法. 中央公論社, 1981.