4J-3

報告書作成基準支援システムにおける文解析について

市吉伸行　岡沢幸一　藤山一敏　和田仁
(㈱三菱総合研究所　応用システム部)

1．はじめに

我々は、客先に納入すべき報告書が、定められた基準を満たしていることを確認する、パソコン上の文書処理システムを設計している。ここではその一環として、現在の開発中の文解析プログラムについて、設計方針と評価結果を述べる。

2．報告書作成基準支援システム

当社には、客先に納入する報告書を作成する際に注意すべき事項を列挙した「研究調査報告書作成の手引き（「ライティング・ミニマム」）」がある。この手引きでは、報告書の体裁、レイアウト等に関する形式基準、文章の読み易さ、文体等に関する文章基準、研究調査、論理構成等に関する内容基準を定めている。多くの報告書がパソコン上のワープロで入力、編集、校正されるようになった現在、ライティング・ミニマムの一部を同じパソコンを用いて自動化しようという試みが本システムである。当面は、文書基準のうちの文体、かな使い、および計量的な値（文の長さ、漢字の割合等）に関するチェックを目標としている。以下、文章基準チェックに用いる文解析プログラムについて述べる。

3．文解析プログラム

3．1　文解析プログラムへの要請

文解析プログラムへの要請は、①パソコン上で動くこと、②実用的なスピードをもつこと、③完全な構文解析は不要だが、文体や慣用副詞・連体詞を認識できること、である。プロタイプをNEC PC-9801 上のProlog-KABA を用いて作成したが、②、③の要請からDCG,BUP など既存の解析ツールは使わずに、形態素解析と文節解析を融合した新規のパーサ（「ストリーム・パーサ」）を開発した。

3．2　ストリーム・パーサ

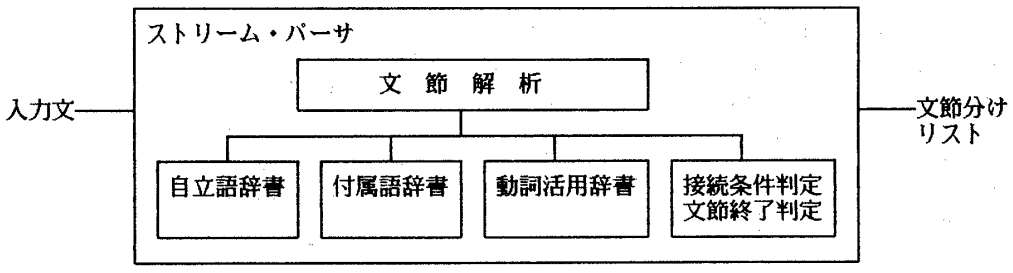
3．2．1　処理

ストリーム・パーサの構成を図１に示す。

ストリーム・パーサの入力は文字ストリーム（リスト）として表された文であり、出力は可能な文節分けの仕方のリスト（全解探索）である。辞書は文字による遷移ネットワークの形で持っている。１文字切出し当りの辞書検索はコンスタント時間である。文節の構造の関する知識は接続関係を検定する情報によって表現されている。接続情報そのものは、Prologの複合項の形で表現されており、ある単語の後接情報と次の単語の前接情報とが統合化可能であることを両単語の接続条件としている。

文に曖昧性がなく、未知単語もない場合、接続条件によるフィルタリングが効くので探索空間は狭く、パース時間はほぼ文字数に比例する。パース実行例を図２に示す。

図１



The Sentence Analyzer of the Writing Minimum Support System
Nobuyuki ICHIYOSHI, Koichi OKAZAWA, Kazutoshi FUJIYAMA, Hitoshi WADA
MITSUBISHI RESEARCH INSTITUTE, INC.

図2

```
?- parse.
1: 彼は広島へいかせられたくなかったらしい。
ambiguity:1
> [代名（彼），助（は）]，
  [未知（体言，広島），助（へ）]，
  [動（五段，行），未然，使役，受身，希望，否定，
   過去，推定]
```

3．2．2 文体の認識

文の属性として文体があり、丁寧調、"だ"調、"である"調、の値をとる（"だ"調と"である"調は排他的でない）。

（例）

説明文も必ずつけます。	…丁寧調
説明文も必ずつける。	…"だ"調，"である"調
図表についても同様だ。	…"だ"調
図表についても同様である。	…"である"調
空が晴れたようです。	…丁寧調
空が晴れましたようです。	…不可（バカ丁寧）

我々のパーサは助動詞のチェックにより文体を認識する。文節構造のチェックがやや厳しいため、上記の最後の例は正しい文と認められない。

3．2．3 かな使い

慣用語のかな使いについては、誤った綴りを辞書に登録しておくことによって綴り誤りを検出する。

3．2．4 未知単語処理

一般に報告書には、専門用語が多用されており、報告書の分野を狭く限定しない限り、辞書にない単語が出現することは避けられない。未知単語が出現する度にパースに失敗するのでは、ほとんど使いものにならないので、未知単語処理が必要となる。我々の文解析プログラムでは、構文解析までは行わないので、未知単語の品詞（および活用形）さえ推測できれば十分である。助詞、助動詞は全てカバーできるので、未知単語は自立語であると仮定してよく、その品詞（および活用形）は接続条件によって推定できる。

未知単語切出しの曖昧性を減らすためのヒューリスティック（主に字種に関するもの）を入れながら、未知単語処理の実験をしている段階である。

4．おわりに

報告書作成基準支援システムにおける文解析プログラムを試作した。汎用機上で推稿や校正を支援するシステムとして、推稿[1],KRITAC[2] などが発表されている。本システムはパソコン上の実用ツールが目標であり、限られたメモリとディスクに収まり、妥当な時間内で処理結果を出さなければならない。

我々は、KWICのようなバッチ的機能は設けず、1文ずつ順に簡単な解析をするアプローチをとった。その際にネックとなるのは文解析プログラムである。試行錯誤的プロトタイピングの必要からPrologを採用したが、現在の段階の評価として、

(1) パソコン上のPrologで処理時間が文字数に比例する文解析プログラムが作成できた。（約10文字／秒）
(2) 未知単語処理については、ヒューリスティックを工夫する事である程度、処理できた。

ストリーム・パーサは、文頭から順に単語を切出して次の単語との接続条件を判定する点で東芝・日英機械システムの形態素レベル解析[3] と類似している。辞書を文字遷移ネットワークで表現したことは、逐次切出し方式と整合性がよく、特に付属語処理の効率を高めている。

今後の課題として、

(1) 自立語辞書の拡張（単語数を増やし、ディスク化する必要がある）
(2) 連語の扱い（独立した文節とみなすかどうか）
(3) 未知単語処理

などがある。(2)，(3) の関連して接続情報の形式を再設計中であり、まず限られた単語数のテキストについて実用試験をしていくつもりである。

参考文献

[1] 牛島，日並，尹，高木：「日本語文章推稿支援ツールのプロトタイピング」，コンピュータソフトウェア，Vol.3,No.1(1986),pp.35-46.

[2] 武田，鈴木，西野，藤崎，丸山：「日本語文書校正支援システムCRITAC」，情報処理学会第32回全国大会，4T-12,1986,pp.1691-1892.

[3] 山中，斎藤，堤，天野：「日英機械翻訳システムにおける日本語解析について（1）」，情報処理学会第32回全国大会，3S-7,1986,pp.1609-1610.