日本語文書校正支援システムCRITACのテキスト・コンパイラ

4J-6

武田 浩一　鈴木 恵美子　藤崎 哲之助
日本アイ・ビー・エム株式会社 サイエンス・インスティチュート

1．まえがき

ワードプロセッサ等により作成された計算機上の日本語文書の校正は、最近になって多数の研究が報告されはじめた分野であり[1-3][8][14]、今後の知的ワードプロセッサあるいはオフィス・システムにとり最も重要な機能であるといえる。これらの研究には対話的校正処理を中心とした校正環境の実現と、いわゆるWriter's Workbench[12]のような各種の文書解析ツールの実現とが含まれる。前者の例としては、EPISTLEシステム[10]のように、実時間の文章解析モジュールと高機能のエディタ（日本語の場合はワードプロセッサ・ソフトウェア）とを結合したものとなり、後者は目的に応じた解析を実行し、結果をファイルに出力するプログラム群[9]からなる。

校正処理を文書作成者の立場から考えると、タイプミス、誤変換等の局所的で自明な誤りはできるだけ文書の作成時に検出されることが望ましく、将来的には仮名漢字変換等の文書入力インターフェイスを改良してこれらを扱うことが予想される。これに反して、文書のスタイルに関するものは、文書入力のリズムを乱すことや、チェックに要するオーバーヘッドを考慮すると、一旦文書作成が済んだ後で一括して処理するほうが好ましい。特に、あるパラグラフ中の文の数が多いとか、同種の語尾が固まって現れるといった大域的な警告をうけても、文書作成者は、かなり前後の文脈を把握してからでないと校正が容易ではない。従って、文書校正処理は、入力時検出と作成時一括検出の２段階で行うのが最適であると判断される。

本報告では、この２段階の文書校正を可能とするために、対話型の文書校正支援システムとテキスト・コンパイラという構成を提案する。テキスト・コンパイラとは、与えられた文書を入力として、校正メッセージと構造化文書[5][14]を出力するソフトウェア・ツールであり、プログラムのコンパイラが、入力となるプログラムから、エラー・メッセージとオブジェクト・コードを出力することのアナロジにより、このように呼んでいる。我々は対話型の文書校正支援システムCRITAC[6][14]を既に試作しており、テキスト・コンパイラはCRITACの文書処理ツールを共有する形で開発中である。以後テキスト・コンパイラをBCRITAC(Batch CRITAC)と呼ぶ。

2．BCRITACの構成

図１にBCRITACの構成を示す。BCRITACは構造化文書上に定義された手続きの集まりである。入力文書を構造化文書に変換するのはCRITACと共用の文書前処理部であり、文書の文節切り、基本語への分割、品詞・読み情報の付加という一連の処理がなされる。基本的な統計情報も同時に計算される。構造化文書が得られると、外部表現や校正知識ベースの呼出しによる校正メッセージの生成が行なわれる。これらの出力は、プリンタ出力用のファイルに含まれる。BCRITACのオブジェクト・コードに相当するものは、構造化文書のほかに、PL/I等の手続き型言語への入力として、

･･･(単語の区切り)<単語の正書>{読み}(単語の区切り)･･･

という可変長レコード形式のファイルを指定できる。次節で述べるように、このフォーマットは簡単に変更できる。

構造化文書はProlog[11]の節集合であるため、Prologで書かれたプログラムからはそのまま処理できる。従って、新たな出力生成プログラムをBCRITAC上で開発・追加することが極めて容易になっている。現在このような拡張性をより高めるために、構造化文書に対する基本操作や文節から導出できる文書構成要素等をPrologで記述しており、これをライブラリとしてCRITACの知識ベースに追加する予定である。

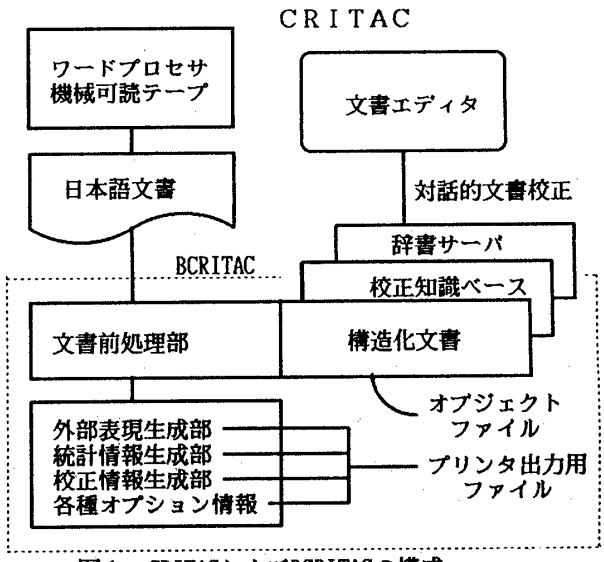

図１　CRITACおよびBCRITACの構成

3．BCRITACの出力例

現在BCRITACの出力には以下のものがある。

[入力文書のソース表現]　ソース表現[6][14]はべた書きの文書表現で、入力文書に最も近い外部表現であるが、文書中の書式情報は取り除かれ、禁則処理等も行なわれていない。BCRITACでは、ソース表現は校正メッセージで指摘される文節を確認するために用いる。

```
                                        左の数字は行番号である
1    1．まえがき
2
3    ワードプロセッサ等により作成された計算機上の日本語文書の校正は、
4    なって多数の研究が報告されはじめた分野であり［1］［2］［3］［
5    14］、今後の知的ワードプロセッサあるいはオフィス・システムにと
6    重要な機能であるといえる。これらの研究には対話的校正処理を中心と
7    正環境の実現と、いわゆるＷｒｉｔｅｒ'ｓ　Ｗｏｒｋｂｅｎｃｈ［１
8    ような各種の文書解析ツールの実現とが含まれる。
9
```

A Text Compiler of a Japanese Text Proofreading System CRITAC

Koichi Takeda, Emiko Suzuki, Tetsunosuke Fujisaki
Science Institute, IBM Japan, Ltd.

[校正メッセージ] 校正メッセージは校正規則に対応した警告文と、その対象となった文節、およびそのソース表現上の行番号と開始桁からなる。ＫＷＩＣ規則による校正メッセージも、ソース表現中の２つの文節に対する警告に翻訳される。

```
* ERROR 31 on line 23 position 6 : 『機械翻訳システム野ためには』
   適当でない漢字が使われています。
   誤変換ではありませんか？

* ERROR 33 on line 40 position 35 : 『実用かと』
   未変換の可能性はありませんか？
   この「か」は漢字で書かれるはずではありませんか？

* ERROR 17 on line 24 position 28 : 『よって』
   １つの文の中に同じ言葉が繰返し使われています。
   他の言い回しができませんか？
```

[入力文書のＫＷＩＣ表現] ＫＷＩＣ表現[6][14]は、文書中の自立語を見出しとして、文書中の各文を重複して並べたものである。並べ方は、読みと正書による順序づけと、見出し語、見出し語に前接する語、後接する語をソート・キーに選択する順序を組合せて指定することができる。標準的な並べ方は、各ＫＷＩＣ行を見出し語、見出し語に後接する語を読みの降順に並べるものであり、このときにはＫＷＩＣ規則[5]による同音異義語の誤変換や表記のゆれの校正メッセージを併せて出力できる。

```
かんようく      、漢字の読み、誤りやすい    慣用句      、言い替
かんじ          をＰｒｏｌｏｇで記述し、    漢字        の読み、
かんじ          もいえるミスタイプや仮名    漢字        変換の誤
かんじ          まり注意しないため、仮名    漢字        変換の誤
かんじ          ）むしろミスタイプや仮名    漢字        変換の誤
かんじひょう    リスト形式に変換し、常用    漢字表      と朝日新
きょくしょてき  名漢字変換の誤りのような  局所的      に現れる
```

[構造化文書] 構造化文書はBCRITACでのオブジェクト・コードに相当し、CRITACによる対話的校正を行なったり、利用者がPrologで記述した特別な処理プログラムを実行するときの入力となる。

```
punc( 4, 1,13,' 、 ').
seg( 4, 1,14,' 将来的には ').
head(4, 1,14,' 将来的 '.nil,' しょうらいてき '.nil,{19},'12S'.nil).
tail(4, 1,14,' に '.' は '.nil,{85}).
seg( 4, 1,15,' 仮名漢字変換等の ').
head(4, 1,15,' 仮名 '.' 漢字 '.' 変換等 '.nil,' かな '.' かんじ '.
```

[書式タッグつき文書] ソース表現では書式情報が取り除かれているため、清書用の文書として開発されているのが本文書である。現在のところ、見出し(:h.)、パラグラフ(:p.)、項目(:i.)、改行(:b.)、図形領域(:f.)の４種類のタッグが作られている。本文書に対するフォーマッタがこれらのタッグと、出力形式の指定から実際の清書ファイルを生成する。

```
15  ワードプロセッサ・ソフトウェア↑）とを↑結合したものとなり
16  目的に↑応じた↑解析を↑実行し、↑結果を↑ファイルに↑出力
17  ラム群↑［９］からなる。↑
18
19 :P
20     校正処理を↑文書作成者の↑立場から↑考えると、↑タイプミ
21  等の↑局所的で↑自明な↑誤りはできるだけ↑文書の↑作成時に
```

オプションとして文節区切記号（↑）を付加している

[統計情報] 統計情報は文書中の非空白文字数、自立語の数、文の数、パラグラフ数、各種の平均値などは文書前処理部で処理される。出現頻度の高い自立語(基本語)や１回しか現われなかった自立語などはＫＷＩＣ表現を生成するときに求めることができる。現在これらの統計値の種類、統計値に基づく文書の均質化[3]やその他の応用[13][15]を検討中である。

[文書の縮約表現] 文書の縮約表現は文頭と文末の一部を、字種の変わり目に従って取り出し、残りを縮退したものである。本表現は各パラグラフの文頭、文末のパターンや、論旨の流れなどを大まかに把握するのに利用できる。

```
１．まえがき

  ワードプロセッサ等により－－校正は、最近になって－－であり［１］［２
］［３］［８］［１４］、今後の－－機能であるといえる。これらの研究－－
実現と、いわゆるＷｒｉｔｅｒ'ｓ－－含まれる。前者の例としては、ＥＰＩ
ＳＴＬＥシステム［１０］のように、実時間の－－結合したものとなり、後者
は－－実行し、結果を－－プログラム群［９］からなる。

  校正処理を－－考えると、タイプミス、誤変換等の－－望ましく、将来的に
は－－予想される。これに反して、文書の－－関するものは、文書入力の－－
乱すことや、チェックに－－考慮すると、一旦文書作成が－－好ましい。特に
、あるパラグラフ中－－多いとか、同種の－－警告をうけても、文書作成者は
、かなり前後－－容易ではない。従って、文書校正処理は、入力時検出と－－
判断される。
```

またBCRITACの起動時に、次の２種類の方法で出力の制御ができる。

[プログラム呼出し] 構造化文書に対して利用者が特別に実行したいPrologのプログラムの呼出しが指定できる。これにより、例えば人名のような特定の語が文書に含まれているとき、その語の読みが間違っていないかどうかチェックし、間違っていれば正しい読みをつけることが可能である。

[プロファイル] プロファイルには上記の各種出力の有無やオプション、文書校正項目の指定(例えば受身のチェックはしない等)を記述する。これにより利用者ごとに出力を調整できる。

４．あとがき

文書校正支援ツールとしてテキスト・コンパイラBCRITACを提案し、現在出力可能な情報について述べた。テキスト・コンパイラにより日本語文書の校正機能の充実と、構造化文書の蓄積による文書データベースを中心としたオフィス・システムの構築が可能となる。

[参考文献]

[1]石井:計算機による日本語の用語・固有名詞の校正、ICOT Rep.,TR-124,1985年7月
[2]牛島、日並、尹、高木:日本語文章推敲支援ツールのプロトタイピング、ｺﾝﾋﾟｭｰﾀ･ｿﾌﾄｳｪｱ、(3,1),1986年1月
[3]絹川:高品質日本語文章作成支援機能の一考察、情報処理学会全大、4H-6,1985年9月
[4]鈴木、武田、藤崎:構造化文書上における校正・推敲手法の検討、情報処理学会全大、4J-5,1986年10月
[5]鈴木、武田、藤崎:日本語文書校正支援システムCRITAC、日本語文書処理研、1986年9月
[6]武田ほか:日本語文書校正支援システムCRITAC、情報処理学会全大、4T-12,1986年3月
[7]武田ほか:日本語文書校正支援システムCRITACの校正知識、情報処理学会全大、4T-13,1986年3月
[8]建石、小野、山田:DictionとStyleの日本語化について、情報処理学会全大、4H-2,1985年9月
[9]Cherry:Writing Tools, IEEE Trans. on Communications,(30,1),Jan.1982
[10]Heidorn et al.:The EPISTLE text-critiquing system,IBM Sys. J.,(21,3),1982
[11]IBM Corp.:VM/Programming in Logic - Program Description/Operations Manual, SH20-6541,Sept.1985
[12]Macdonald et al.:The Writer's Workbench: Computer Aids for Text Analysis, IEEE Trans. on Communications,(30,1),Jan.1982
[13]Misek-Falkoff:Data-Base and Query Systems: New and Simple Ways to Gain Multiple Views of the Patterns in Text,IBM Res. Rep.,RC8769,Mar.1981
[14]Takeda,Fujisaki,Suzuki:CRITAC - A Japanese Text Proofreading System, Proc. COLING'86,Aug.1986
[15]Tankard:The Literary Detective,BYTE,(11,2),Feb.1986