日本語文書校正支援システムＦｌｅＣＳの誤り検出処理

3J-4

清水　周一
日本アイ・ビー・エム株式会社　東京基礎研究所

1．はじめに

日本語文書中の誤りを計算機により指摘する、校正支援システムの研究は、これまで多く行なわれてきている[1,2,3]。その一手法として、漢字かな混じり文を形態素レベルで解析し、解析用辞書中の単語では接続が不可能となる部分を、未知語として指摘する方法が多くのシステムで採用されている。しかし、日本語ワープロで作成された文書に対しては、英文におけるスペル・チェッカーほどの効果はない。漢字かな混じり文の入力方式として、かな漢字変換方式を用いているため、例えば、「欠裂(決裂)」や「速時(即時)」といったような書き間違いは起こらない。文書の修正時に、削り過ぎや削り残し等のため若干生じる程度である。それどころか、タイプミスした入力にさえ、形態素の接続関係の満たされた結果を出力しようとする。このような、形態素の接続では判定ができない誤りを、高いヒット率で、しかも高速に検出する実用的な校正支援システムを構築するには、ヒューリスティックな誤り検出ルールを数多く備えることが最も現実的であると考えられる。

われわれは、実用的な校正支援システムを構築するためのツールとして、ＰＣのＯＳ/２上にＦｌｅＣＳ (Flexible rule-based Critiquing System) を開発してきた。ＦｌｅＣＳでは、校正ルール（誤り検出ルール）のために、柔軟なパターン記述用言語を提供している。

2．誤りの分類と校正処理の対象

日本語ワードプロセッサで作成された文書中によく現われる誤りで、キー操作に関連しているものは、以下のように分類できる。

(A1) 入力ミス
　例.「そけで(そこで)」、「スッタク(スタック)」
(A2) 修正ミス
　削り過ぎ、削り残し、など
(A3) 変換ミス
　同音異義語の選択誤り、表記・送り仮名の揺らぎ、変換し過ぎ、など

原文、あるいは頭に思い浮かんだ文章そのものに誤りが含まれる場合、以下のような症状が見られる。

(B1) 重言
　例.「再婚しなおす」、「第一日目」
(B2) 文体の不統一
　「だ/である」調と「です/ます」調など
(B3) 単語のレベルの不統一
(B4) 構文的な誤り
　主語と述語の格が合わない、など
(B5) 意味の曖昧さ
　文節の係り受け関係の曖昧さ、代名詞の曖昧さ、など
(B6) 読みにくい
(B7) 内容自体の誤り

文法的には問題がないが、業種などによって特に制約が設けられているために、誤りとなるものには次のような項目がある。

(C1) 禁止語
　例.「盲」
(C2) 経済用語の送り仮名規則
　例.「受付け期間(受付期間)」
(C3) 第一水準以外の漢字の使用
　例.「語彙(語い)」

われわれは、見かけ上単純なため、校正・推敲者にとって見逃しやすいと思われるＡ群、Ｃ群の誤りを計算機で検出することを目的としている。Ｂ群の項目のような校閲・推敲対象は、単純な重言などを除いて、扱わない。これらの項目は、校正・推敲者にとって比較的見つけやすいものであること、また、文書作成の段階で、注意深く単語が選ばれ文が練られていると考えられること、などのためである。

3．誤り検出ルール

文書に対して形態素解析を行なうことにより、局所的な文法の誤りはある程度検出できる。しかし、局所的に誤りを含むにもかかわらず、形態素の接続関係が満たされている場合も多い。例えば、ワープロ文書で「は」や「が」などの助詞を誤って変換し、「歯」や「画」とした場合でも、文法的な誤りにならないこと

Error Detection in Japanese Text Critiquing System FleCS
Shuichi SHIMIZU
IBM Research, Tokyo Research Laboratory

が多い。われわれは、このような誤りを検出するために、正規表現パターン記述を用いて、ヒューリスティックな校正用ルールを実現している。また、表記の揺らぎなどのような、正規表現では記述できない誤りパターンを検出するために、柔軟な校正ルール記述用言語を用意している。

正規表現パターンは、一文節内の形態素の列に対して記述する。パターンの単位は、文字ではなく、形態素である。したがって、パターンの単位の条件部には、形態素の様々な属性が記述できる。また、品詞や文字列の他に、それらをキーとして特殊辞書を検索する、といった強力な条件を柔軟に記述することができる。

```
%<words:  [POS($名詞)&&DICT($形容動詞)]
          [POS($名詞)&&STRING("名")] >%
```

「優雅-で/自由-名-生活-を/楽し-む-。」

図1 過変換誤りを検出する正規表現パターン

ここで、図1に、過変換を検出するための例を示す。一般に、ユーザは、かな漢字変換処理を促すキー(変換キー)を押す場合、必ずしも文節の区切りで行なっているとは限らない。特に、長い複合名詞などを入力した直後で、助詞をタイプする前に変換キーを押すことがある。図に示した過変換誤りは、「...でじゆう」とタイプした後、変換キーを押し、続けて「なせいか...」と入力したときに起こる。この誤り文を形態素解析しても、「自由名生活」の部分は、複合名詞であると解釈されるため、誤りは検出されない。そこで、われわれは、誤り文に対して形態素解析した結果を抽象化し、正規表現で記述することによって、このような誤りを検出する方法を採る。図の例では、ある条件を満たす二つの形態素の並びを検出する。まず、一つ目の形態素は、品詞(POS)が名詞であり、その表記をキーにして辞書を参照する(DICT)と、接続によっては形容動詞にもなりうるものである。二つ目は、品詞(POS)が名詞であり、表記(STRING)が「名」である形態素である。この正規表現パターンにより、「..自由名..」や「..平和名..」などが検出できる。

文節の範囲を越える誤りパターンの記述例を図2に示す。この例は、「いなか」と「田舎」、「取扱い」と「取り扱い」のような、名詞の表記の揺らぎを検出するための校正ルールである。ルールの条件部は、条件要素(図中の{...})の連接として記述される。条件要素は、文節内の評価だけで完結する単項条件と、文節間の性質を評価する2項条件により記述される。例では、"isNoun()"、"isHonsoku()"、"isKyoyou()"が単項条件、"String(@p1.自立語)=Seisyo(@.自立語)"が2項条件である。前述の正規表現パターンは、単項条件として扱う。図の例の条件部の読み方は、次のとおりである。まず、文節(PHR)があり、その自立語部(@.自立語)は名詞で(isNoun)、その表記は「本則」に従う(isHonsoku)。また、別の文節があり、その自立語部は名詞で、その表記は「許容」の範囲である(isKyoyou)。そのような二つの文節について、両方が同じ正書を持つ(sameSeisyo)ならば、この条件部は満たされる。ルールの実行部(+{...}+)では、条件部を満たした文節に対して、誤りのメッセージなどを付加する。

4. おわりに

実用的な日本語文書校正支援システムを構築するために、われわれは、ヒューリスティックな校正ルールを柔軟に記述できる枠組みを用意した。今後は、このようなルールを用いた校正処理について、検出不足や誤検出などの評価を行なう予定である。

[参考文献]

[1] 武田他：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(1986)4T-12
[2] 安田他：日本語訂正支援システムREVISE. 情報処理学会第33回全国大会(1986)4J-9
[3] 小山他：文章校正支援機能における日本語解析. 情報処理学会研究報告(1988)88-NL-69-2

```
@p1:PHR { isNoun(@.自立語); isHonsoku(@.自立語) }
@p2:PHR { isNoun(@.自立語); isKyoyou(@.自立語);
                String(@p1.自立語)=Seisyo(@.自立語) }
+{
    setWarning(@p1,"名詞の表記が揺れています(本則 <->「%s」)",@p2.自立語->string);
    setWarning(@p2,"名詞の表記が揺れています(「%s」<-> 許容)",@p1.自立語->string);
}+
```

図2 名詞の表記の揺らぎを検出するルール