文章推敲開発支援システム

3S－8　杉江美佐子　雷海涛　齋藤啓司　斎藤裕美
(株)東芝 マルチメディア技術研究所

1　はじめに

我々は従来より、文章の推敲を目的とするシステムの開発を行なってきている。「文章推敲システム」の開発には、解析精度の向上、誤り検出の精度の向上などが重要であるが、解析規則、誤り検出規則の量が多くなると、相互に関係している規則なども増え、規則の変更時に慎重な検査が必要になる。

また、商品化した場合はどのようになるか、ウインドウやマウスによるMMIを持つツールを使って検討したいという要求もある。しかも、「文章推敲システム」の仕様に変更が生じた場合にも、ツールを変更することなく、新しい「文章推敲システム」の仕様に簡単に合わせられるツールにしておきたい。

そこで我々は、規則変更時に必要な検査を自動化するツールである「正解率測定ツール」と、ウインドウ環境で「文章推敲システム」の評価を行なえる「対話的評価ツール」の2種類のツールからなる「文章推敲開発支援システム」を開発した。このシステムは、「文章推敲」の機能強化などに柔軟に対応できるように、データのやりとりはテキスト形式で行なうようにした。

2　検査の自動化

「文章推敲開発支援システム」と「文章推敲」の関係を図1に示す。斜線の部分が、「文章推敲開発支援システム」である。便宜上「正解率測定ツール」と「対話的評価ツール」をあわせて「推敲ツール」と呼ぶことにする。「文章推敲」と「推敲ツール」間のデータのやりとりは、テキスト形式のある書式に従ったものを仲介して行なわれる。また、「推敲ツール」の動きは「環境ファイル」によって制御されている。

検査をする場合には、ある文章を推敲したらどのような誤りを検出できるかを記述してある「正解ファイル」をあらかじめ作成しておく。そして、その「正解ファイル」と「文章推敲」がテキスト形式で出力した結果を比較する。

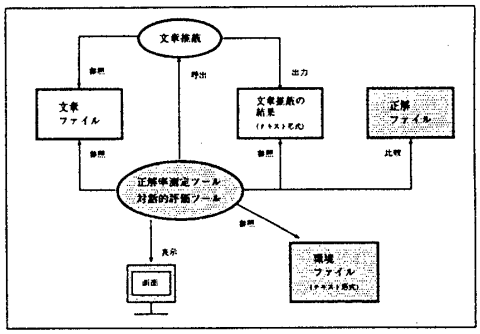

図1:「文章推敲開発支援システム」と「文章推敲」の関係

「文章推敲」で検出できる誤りにはいろいろな種類があり、異なる種類の誤りが同一の箇所から検出されることもある。しかもその誤りがそれぞれ推奨候補を持つ場合には、検出したそれぞれの誤りを修正するのに満足する推奨候補を提示する必要がある。また、個別の誤りの修正にも対応する必要がある。そこで、同一箇所で複数の誤りを検出した場合にはそれらをひとつのグループとして扱い、その中で誤りを組み合わせ、その組み合わせの誤りを修正するのに満足する推奨候補を提示するようにした。

例文を用いて説明する。「私は久しぶりに手紙を描きました。」という文があった場合、「文章推敲」は図2のような結果を出力する。1行目の「$」は、このファイルが「文章推敲」の出力であることや、いつ出力されたものであるかの情報である。この文では「描きました」という部分で「同音語誤り」と「文体(である体)」の2種類の誤りが検出され、「同音語誤り」と「文体(である体)」、「同音語誤り」、「文体(である体)」の組み合わせが考えられる。従って、グループを示す「<G>」と「</G>」の間に3種類の組み合わせを記述する。なおここでは英記号が処理の単位になっていて、「/」のつくものまでがその対象になる。

Text Polish Development Support System
Misako SUGIE Haitao LEI Keiji SAITOU Hiroyoshi SAITOU
TOSHIBA CORPORATION MULTIMEDIA ENGINEERING LABORATORY

組み合わせは「S」で記述する。ただし、組み合わせによっては対象箇所が同一にならないものも出てくるので、同一箇所の誤りを「C」で示すことにした。つまり、検出した誤りの最小単位が「C」から「/C」になる。そこには、「誤りの種類」「位置」「置換候補」「メッセージ文字列」が「C」「P」「<K>」「<M>」で示す部分に記述される。

もちろん、単独の誤りを検出した場合には、「<G>」ではなく「C」からが記述される。

```
<$ form=suikou user=suikou create=Fri Jul 22 11:45:08 1994>
<G>
  <N>
    原文「私は久しぶりに手紙を描きました。」
  </N>
  <S select= 同音語誤り; 文体 (である体)>
  <C type= 同音語誤り; 文体 (である体)>
  <P start=1,8 end=1,10;start=1,11 end=1,16>
  <K>
    手紙を書いた。
  </K>
  <M>
    [推奨候補]
      手紙を書いた。
  </M>
```

図 2: 「文章推敲」出力結果の一部

「文章推敲」の出力と「正解ファイル」を比較した結果は、図3のようになる。はじめに、テキストファイル名や正解ファイル名などの情報を表示し、もし正解ファイルと異なる部分があったら、その次に異なっていた誤りが何であったかの情報が出る。この場合には、「文体(である体)」と「同音語誤り」に関する不具合のあることがわかる。開発者は該当する部分を修正し再度検査を繰り返す。

3 「文章推敲システム」の試作環境

ウインドウ環境で「文章推敲」の試作評価を行なえる「対話的評価ツール」の画面を図4に示す。このツールの特徴は、テキスト形式の環境ファイルと呼ぶものによりツールの動作が制御されている点である。例えば、この画面にあるメニューや「文章推敲」のコマンド名、起動時のオプションなどは環境ファイルに登録してあるものを使う。そのため、「文章推敲」の仕様に変更が生じ、たとえば検出できる誤りの種類を増やした場合にも、環境ファイルにツールのメニューに表示する名前や「文章推敲」にその誤りの検出をさせるためのオプション名を登録をすることにより、ツールのプログラムを変更することなく使えるようになる。

また、環境ファイルにはウインドウ環境での色や下線などの表示環境や誤りを検出した場合のアクションの指定などもできる。

4 おわりに

文章の推敲を目的とするシステムの開発を支援するシステムについて報告した。文章推敲システムの開発には、解析精度の向上、誤り検出の精度の向上などが重要である。解析精度の向上、誤り検出の精度の向上のためには、解析規則、誤り検出規則を充実させていくことが必要であるが、規則が増えるとそれだけ規則の変更が難しくなり、変更時には慎重な検査が必要になる。本システムは、正解ファイルと比較することにより検査を自動化できる。

また、「文章推敲システム」の開発では、マンマシンインターフェースの検討も重要であるが、本システムでは、環境ファイルを用いたウインドウ環境のツールにより、商品化時と同じような環境を手軽に実現することができる。

同音語誤り／文体 (である体) :
(25,8) 手紙を書いした。

	正解率	チェック余計	チェック洩れ
文体 (である体)	95%	0 件	1 件
同音語誤り	88%	1 件	2 件
語の順番	100%	0 件	0 件
対になる語の誤り	100%	0 件	0 件
語の重複	100%	0 件	0 件

図 3: 「正解率測定ツール」出力の一部

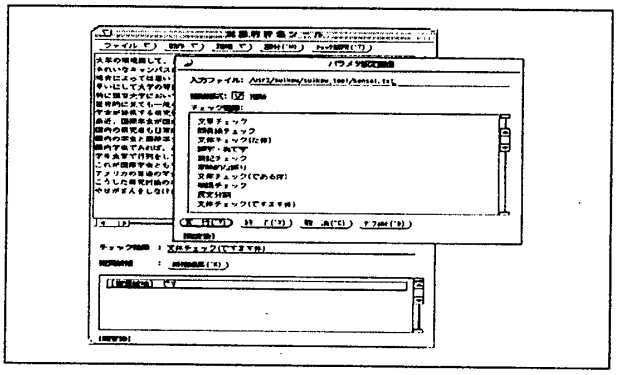

図 4: 「対話的評価ツール」の画面

参考文献

[1] 構文解析を用いた文章推敲支援システム
杉江美佐子、後藤浩文、中里茂美、大黒和夫
情報処理学会第４１回 (後期) 全国大会