4J-12

日本語文の誤り検出に関する研究

山田洋志　内田幸司　荒木健治　栃内香次　永田邦一
北海道大学工学部

1．まえがき

近年，計算機を用いて文章を作成することが次第に多くなっている．しかし，これらのほとんどは単にかなを漢字に変換するために計算機を利用しているだけで，文章の推敲を補助してくれるものはあまりない．文章の誤りには様々な種類があり，綴りや送りがなの間違いのような単純なものから，記述された内容の間違いまで広い範囲に及んでいる．本稿では，文章の意味内容の深い理解を必要とせずに見分けられるかかり受けの誤りに着目し，それをいくつかのパターンに分類し，検出の方法について考察する．

2．誤りの分類と実例

昭和59，60年度の電気関係学会北海道支部連合大会講演論文集のうち182篇を調査し23文を誤り例として取り出した．誤りの原因として多かったのは，4つのパターンであり，それらで17例を占めていた．以下にそれらの内容と実例，誤りの原因となった場所を示す．

1)　書き出しと結びの不一致

文章の書き出しと結びの部分がうまく対応していないもの．

例　本プログラムはC言語の特長を生かし，（中略）回路の変化にも即対応できるようにした点である．

この例では始めの部分を「本プログラムの長所は」などとするとよい．

2)　助詞の誤り

助詞の選び方がおかしいもの．単なる書き間違いと思われるものも多い．

例　これを5kHzのLPFを通し，12kHz，12bitで量子化した．

「LPFを」を「LPFに」にかえる．

3)　受動，能動，使役の誤り

「する」と「される」，「する」と「させる」の使い分けが不適切なもの．2)の場合に含めることのできるものもある．

例　InP基盤の場合には，最初にGaAsをバッファ層として成長させた後，その上に超格子を成長した．

文末を「成長させる」にする．

4)　言葉の重複

同じ語が必要以上に繰り返して出てくるもの．これは必ずしも誤りではないが，推敲の対象になると考え誤りに含めた．

例　文献の自動抄録に関する研究は，アメリカの情報処理学者P.H.Luhnがキーワード密度法による自動抄録のアルゴリズムを提唱して以来さまざまな抄録法が研究されている．

文頭を「文献の自動抄録に関しては」などに変える．

いずれのパターンについても誤りであるかどうかは断言できず，個人の好みの問題となるものも多い．

学術論文などでは，ページ数に制限があるため，長い挿入句や，複雑な修飾関係が多く用いられ，上記のような誤りを見つけることを難しくしている．

Detection of Error in Japanese Sentences
Hiroshi Yamada, Kouji Uchida, Kenji Araki, Koji Tochinai, Kuniichi Nagata
Hokkaido University

3．修飾の削除による推敲の支援

前節のような誤りの中には，文中の挿入句や修飾語とは無関係に，対応する語の関係だけで見つけられるものが多い．そこで，文から，主語，目的語，述語の部分を取り出して表示し，推敲の手助けとすることが考えられる．以下にそのアルゴリズムの概要を示す．

1)　文の分割

原則としてひらがな列の終わりで分割する．ひらがな語などは例外として辞書に登録しておく．複合語の分割は行わない．

2)　文節の分類

各文節の末尾を調べ，助詞（て，に，を，は，が，も，や），動詞・形容詞・形容動詞の連体形・連用形，副詞に分類する．まぎらわしいものは辞書に登録しておく．

3)　文節中の動詞の検出

2)の文節末の分類で動詞と判断されたものの他に，サ変動詞の活用語尾の検出，助動詞「られる」の活用形の検出，動詞を受ける副詞（ために，ように）の検出を行い，文節中に動詞が含まれているかを判定する．

4)　文中のかかり受け関係の決定

助詞「は」，「が」，「を」，「に」で終わる文節について，どの文節にかかるかを決定する．「は」で終わる文節は文の最後の動詞，その他については直後に現れた動詞にかかるとする．ただし現在のところ１つの文節が複数の文節にかかる場合は考えていない．

以下に実際の例を示す．

計算機は数値を計算する機械として誕生した．

```
  zyo  ha   計算機は
  zyo  wo   数値を
D dou  tai  計算する
  zyo  te   機械として
D end  dou  誕生した．

No. 1  数値を　計算する
No. 2  計算機は　誕生した．
```

２節で述べた誤りの実例について上記のアルゴリズムを適用し，修飾を取り除いた文から誤りが読み取れるかを判定したところ，17例中７例については誤りを確認できた．

4．誤りの自動検出

２節の1)については，誤りかどうかが文全体の構造や内容に左右されやすく，完全な自動検出は難しい．しかし，この誤りは比較的長い文で起こる傾向があるので，前節の方法で挿入句などを取り除くことで人間による発見が容易になる．

2)については各動詞について受けることのできる格の種類を登録したデータベースを作成すれば誤りを検出できる．

3)についても，名詞と動詞をそれぞれ分類して，主語になりうるかどうかの情報を登録すれば検出することができるが，2)の場合以上に巨大なデータベースが必要になる．

4)は文中の語の頻度を数えることで容易に検出できる．ただし，実際には誤りでないものの含まれることがある．これについてはプログラムを作成して実験を行いいくつかの文が選び出された．以下に例を示す．実験用データには情報処理関係の論文を用いた．

例　・この動作は普通のスタックで可能な動作である．

・よって黒い要素が最終的な要素であり，白い要素は背景であると決定する．

始めの例については２つの「動作」のどちらかを削った方が良いが，２番目の例についてはそれぞれの好みの問題であろう．

5．おわりに

文章のかかり受け関係の部分に原因のある誤りを分類し，文の一部を取り出しても誤りが見つけられることを示し，さらに自動検出の方法についても述べた．今後はアルゴリズムの改良を行いつつ自動検出の実験を行う．