# 単語の概念を採り入れた日本文チェックの一方式

5P－5

下村 秀樹　　高橋 延匡
(東京農工大学　工学部　電子情報工学科)

## 1．はじめに

ワードプロセッサなどの普及により，計算機で日本語の文書を扱うことが一般的になってきた．しかし，文章中の間違いのチェックは，まだ人間が行っているのが実状である．この点に関して，ユーザからの「タイプミスや変換ミスから生じる単純なミス程度のチェックは，計算機で行ってほしい」という要求は非常に大きい．

英文では，単純な綴り間違いを指摘するツールとして，スペルチェッカがある．これは，単語ごとの辞書検索を基本とした単純な処理でありながら，有用なツールとして利用されている．一方，日本語では単語ごとに分かち書きをする習慣がなく，単語レベルの処理が難しい．そのため，文字コードだけを利用した字面処理が，文チェックの一つの有効な方法として提案されている [1]．

我々はこれを一歩進めて，単語の概念を取り入れた文のミスチェックの可能性の研究を行っている．本稿では，単語の概念を利用した文チェックの可能性と，その処理の概要について述べる．

## 2．チェック対象のミス

日本文の単語を認識するためには，形態素解析が必要である．しかし，形態素解析で文を単語に分解しただけでは，文の意味に関するような，高度なチェック処理はできない．このレベルでは，文にならないような，おかしな単語や表記のチェック程度が限界であると考える．具体的には，キーボードで入力された文章の，キータッチミスによって起こる変換ミス，文字入力ミスをチェックの対象とする．

## 3．形態素解析のコストを利用したミスチェック

### 3.1 コスト付きの形態素解析

一般に形態素解析はグラフ構造に表現することができ，その処理はグラフの最小コストのパスを捜す問題に帰着する [2]．我々は，探索のコストとして，

(1) 連続する単語の品詞間の接続確率
(2) 品詞の表記の確率

の二つの確率を用いた，形態素解析を実現した．(1) は，文法を「名詞の後は助詞が多い」などように確率的に表現したものであり，(2) は，「名詞は平仮名では書かれにくい」など，文の表記の特徴を表現している．実際の解析処理では，この確率の－logをとり，それぞれ，「接続コスト」，「表記コスト」として各単語に与えた．

この形態素解析の正解率は，現在約95%（解析率 ＝ 正解文字数 ／ 全文字数）である．

### 3.2 形態素解析のコストの分布

文中にキータッチのミス（タイプミス，変換ミス）がある場合，さきに述べた形態素解析結果に次のような現象が起こると予想される．

(1) タイプミスによる誤字，脱字などの場合
・解析に失敗する
・おかしな単語に解釈され，単語のコストが大きくなる
(2) 変換し忘れた場合
・単語の表記コストが大きくなる
(3) 変換ミスをした場合
・違う品詞に変換された場合，接続コストが大きくなる

つまり，解析結果の各単語に付けられたコストに注目すれば，ミスが指摘できる．

この可能性を検証するために，ミスを含む文とミスのない文を解析して，ミスのある部分とミスのない部分の解析結果の単語のコスト分布を調べた．ミスがあるかどうかは，数人に文章をチェックしてもらい，人間がミスと判断したものからタイプミス，変換ミスと見られるものを抜きだした．結果を図1に示す．

ミスのない部分の単語のコスト分布（図1a）は，表記コスト，接続コストともに2.5以上のところでほぼ0になるような，急な減衰を見せている．コストが2.5以上の単語は表記コストで2.8%，接続コストでは1.4%である．

それに対し，ミスがある部分のコストの場合，コスト3以上のところにもかなり多く分布している（図1b）．これは，ミスのない単語のコスト分布とは明らかに異なる分布を示している．基準値の取り方にもよるが，この方法で十分にミス抽出ができると考えられる．

## 4．日本文ミスチェックツール

### 4.1 単語リストを利用したミスチェック

3章での実験をもとに，日本文ミスチェックツールを作成した．ミスチェックの基本的な方針は，「形態素解析結果のコストが基準値以上の単語をユーザに知らせる」と

---

A Method for Checking Japanese Sentence with Concept of Word
Hideki SHIMOMURA, Nobumasa TAKAHASHI
Department of Information Science, Faculty of Technology,
Tokyo University of Agriculture and Technology

いうものである．しかしこれだけでは，変換ミスの結果が同一品詞の別の単語に変換された場合などには全くチェックできない．

これに対して，「単語リスト」を利用したチェックをユーザが行う方法を考えた．「単語リスト」とは，文章中に出現した単語とその出現回数と出現位置のリストである．単純なミスは繰り返して，数多くは起こらないと仮定すれば，ミスを含む単語の出現回数は少ないはずである．そこで，単語リストを出現回数の少ない順にソートしてユーザにみせることにより，ユーザが間違いを発見できる．形態素解析により単語の概念がはっきりしているため，単語リストの生成は容易である．

## 4.2 ユーザインターフェイス

本チェックツールは，図2に示すような2つのモードを持つ．

一つは文書の内容をそのまま表示し，ミスの可能性のある部分（コストの高い単語）にアンダーラインを引き，ユーザに知らせるものである（図2a）．アンダーラインの部分をマウスでクリックするとチェックした理由が下のウィンドウに表示される．

もう一つの画面は，単語リストである（図2b）．ここでは，出現した単語が出現回数の少ない順に並び，各単語の出現位置が単語に右側に表示される．この画面でユーザが単語の出現位置をマウスでクリックすると，下のウィンドウにその単語の付近の文が示され，KWICとなる．

## 6．おわりに

本稿では，単語の概念を利用した日本文チェックとその処理概要について述べた．今後も自然言語処理や，日本語文章のミスのチェックなどについての研究を行い，最終的には文書作成支援システムに発展させて行きたいと考えている．


## 謝辞

本研究の形態素解析で使用した日本語辞書の原データである「九州芸工大自立語辞書KID－J82」を提供してくださった，九州工業大学の吉田将教授に感謝する．


## 参考文献


[1] 牛島他：日本語文章推敲支援ツール「推敲」の使用について，九州大学大型計算機センター広報，Vol．18，No．1，pp．35－46，(1986)

[2] 吉村他：コスト最小法を用いた日本語文の形態素解析，情報処理学会自然言語処理研究会報告 60－1，(1987)


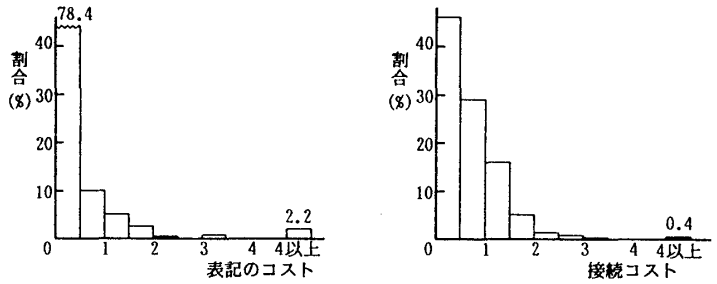


図1a　ミスのない単語のコスト分布

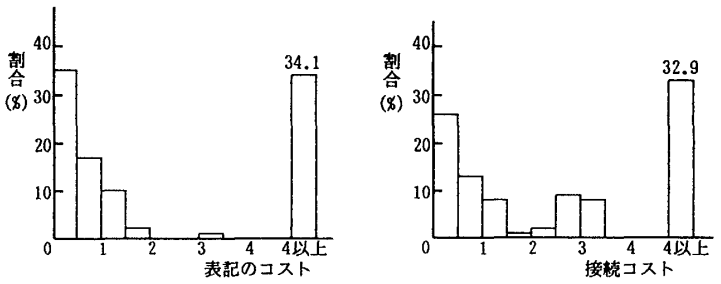


図1b　ミスのある単語のコスト分布

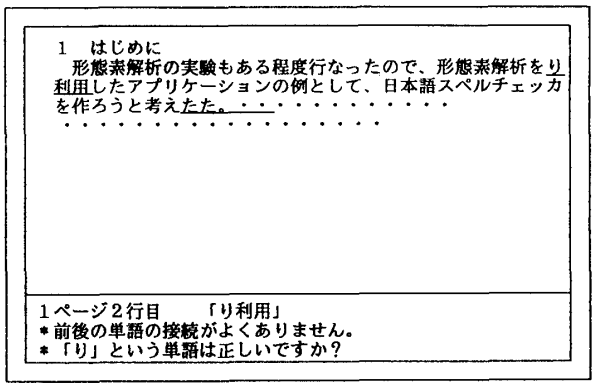


図2a　形態素解析のコストを利用したチェックの画面

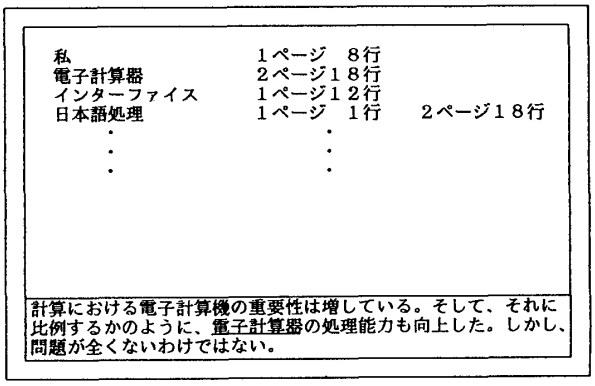


図2b　単語リストを利用したチェックの画面