自然言語処理 117－21
（1997. 1. 21）

校正支援システム Joyner における表記誤りの訂正方式

伊吹 潤　徐 国偉　斉藤 孝広　松井 くにお
{ibuki,guowei,takahiro,kunio}@ling.flab.fujitsu.co.jp
富士通研究所

〒211 川崎市中原区上小田中 1015

従来の日本語の校正支援システムでは様々な手がかりをそのままユーザに提示しており，情報の信頼性や誤り可能性の判断の大部分をユーザが行なう必要があった．我々はこれに対して誤り仮説生成部と検証部を独立させた日本語の表記誤りの自動訂正のための新たな枠組を提案する．

本構成によって，システムで必要な様々なデータを独立に管理してテキスト分野移行や様々な入力手段への対応の容易さを実現し、又辞書データによる検証を経たデータのみを提示することによって広い範囲の表記誤りに対して信頼性の高い情報を提供している。

A new approach for Japanese Spelling Correction

Jun IBUKI　Guowei XU　Takahiro SAITOH　Kunio MATSUI

Fujitsu Laboratories Ltd.

1015, Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, 211 Japan

Although several tools do exist for the detection and correction of Japanese orthodoxical errors, they either deal with too small part of the whole range of errors, or fail to provide reliable error information.

We propose a new system for Japanese error correction, which consists of two indepently functioning parts : hypothesis generator and verificator.

Hypothesis generator detects possible orthodoxical errors and assumes their original spelling from the input text, while the verificator looks up basic dictionary and word-to-word co-occurence relation to sift out improper hypotheses.

1 はじめに

我々はテキストの作成や再利用の過程で問題となるテキスト中の表記の揺れの統一や単語表記の誤りの検出訂正のシステムの研究開発を行なっている.我々は新聞校正の現場での専門家や,個人的なワープロ・ユーザ等のための校正支援システムの研究を同時に行なっており,そのため様々なユーザやテキストの分野に容易に対応できることが不可欠となっている.

また新聞校正の現場では校正作業の効率化という観点からできる限り確度の高い誤り情報を提供することが求められ,我々はこうした要望をとり入れ,新たな日本語の表記誤りの校正支援システムの新たな枠組を構築している.

第2章でまず表記誤りの種類や原因について考察し,第3章で既存のシステムがどのように前出の誤りを処理しているかについて見る.それ以後の章では我々の技術的な目的とそのためにとったシステム構成と評価結果について述べる.

2 表記誤りの種類

テキスト中に現われる表記誤りの種類と頻度はテキストの入力方式に依存する部分が大きい.一般のワープロでは仮名漢字変換のプロセッサを用いてキーボードを操作して入力するが,新聞作成の現場ではタブレットによる入力を含む複数の種類の日本語入力方式が存在しているために,こうした入力に付随する誤りも考慮の対象とすることとした.

我々は誤りの訂正を辞書情報との照合を基本として行なう戦略をとっており,これにはまず単語列の認識(形態素解析)が成功するかが重要なポイントとなる.ここでは表記誤り全体を形態素解析が成功するかによって,2種類に大きく分けてその中を更に考察した.

2.1 単語レベルの誤り

漢字の同音異義語による誤り

直接の原因は仮名漢字変換における誤変換であるが,その要因からユーザが意図して誤ったかどうかでいわゆるケアレス・ミスと認知誤りに分類できる(以後,誤り例は括弧内に正しい表記を付けて示す).

- 認知誤り

  ユーザが対象とする語彙や使い分けの知識がなかったり,誤って憶えているために起きる誤り.主なものとして次の2つが挙げられる.

  - 意味的に類似性が高く混同しやすい誤群間の使い分け

    ex. 犯人の追究(追及)

  - 固有名詞(チェックに実世界の知識が必要)

    ex. 村山富一首相(橋本龍太郎首相又は村山富市元首相) 国鉄精算事業団(国鉄清算事業団)

- ケアレス・ミス

  変換対象の文字は正しく入力されても,ユーザが変換結果の確認を怠っていた場合に起きる.仮名漢字変換の方式や辞書に依存する部分が大きい.

  ex. 私自信(私自身) 家へ買える(家へ帰る)

漢字の類義語による誤り

英語の学習者が不規則変化動詞の活用を間違えるのと同様語彙に関する知識が不足して,既に対応する単語があるにもかかわらず造語規則によって新たな単語を作り出してしまうもの.

ex. 非安定(不安定)

2.2 文字レベルの誤り

ミスタイプやある文字を別の文字として誤用することによるもの.

漢字の同音文字,類形文字による誤り

これらは文字単位の入力をした際に同じような漢字と混同した結果と思われる.ワープロ入力の場合は読みの難しい人名の場合等にこのような誤りが散見されるが,日本語タイプによる入力の場合はむしろこちらの誤りの方が割合が多くなる.

ex. 不情理(不条理),均衡(均衡)

カタカナ語句における表記の揺れ,誤り

音韻の構造の違う外国語の単語は音表文字に移すとしても様々な表記方法が存在して,表記の揺れの大きな原因となっている.又カタカナ語を入力する場合には変換をかけずに字種を指定してそのままタイプする場合が多く,辞書の語彙チェックがかからないためにミスタイプがそのまま残る確率が高い.

ex. ショウ(ショー),人間ドッグ(人間ドック)

平仮名単語の表記の揺れ

二重母音に対する表記方法,送りがなの揺れ等がある.量として多いのは送りがなの揺れによるものである.送りがなの場合については一般に許容されている範囲で複数の表記が存在するが,公用文基準や文部省基準等に統一する場合や,同一テキスト中の表記の揺れあれば,どちらかへ統一することが求められる場合等がある.

ex. そのとうり(そのとおり),行なう(行う)

漢字表記と仮名表記の揺れ

常用外漢字の使用を認めるかどうか, 接続詞（ex. 又, 猶）等を漢字で書くか平仮名表記とするか等の揺れに基づくものである.

こういった表記基準については基準が新聞社間で逆になっていたりして, 校正基準にぶれがある. このことから人間が基準をどちらかを知らずに誤ることが多い.

3 既存のシステムでの対処

ここでは既存のワープロや校正支援システムを中心として前節で挙げた誤りをどう対処しているか, どのような問題があるかについて述べる. ここでは特に誤りの検出の精度, 同種の誤りのどの程度をカバーしているか, 誤りをどう訂正するかについての情報をもつかについて検討する.

校正用単語の登録 自動訂正技術としては, 単語の誤った表記のパターンを予め辞書に登録しておき, 形態素解析の結果, その単語が出現した時に正解の単語を指示するものが一般的であり, 広く実用化されている [1]. 例えば「安全補償」という誤り語を「安全保障」という正解語情報と共に辞書に登録して置き, 解析結果中にこのような単語があれば, 正解と共に提示する訳である.

このような仕組みは単純ではあるが, 送りがなの揺れ等のよく起きる誤りのパターンがある程度限られているような場合は有効であり, 誤り指摘の適合率の点でも優れている（本来の誤り以外のものが指摘される可能性が低い）.

しかしデータベース中に誤りの種類に関する知識（保障を保証と間違える）が語彙（安全保障という言葉）に関する知識と不可分の形で保持される訳であり, テキストの分野や入力に対して独立してデータの調整をすることが不可能である点, 或は特に文字レベルの揺れをとり込むと辞書の語彙数が爆発するといった点で問題がある.

誤り易い同音異義語情報の提示 仮名漢字変換の誤り検出に対しては, ある単語に対して混同しやすい同音語を使い分けの基準と共に提示し, ユーザ自身がチェックをする手助けとする仕組みがある（更に共起する言葉によって正誤の判断をすることもある）.

ただし, システムで登録されたグループは一般的な語彙であり, かつ意味的に似ていて混同しやすいもののみである. 従ってケアレス・ミスは対象とはならない.

一方, 指摘が有用な情報であるかはテキストの作成者にとって新規の知識であるかに依存する. 自分の知っている言葉の使い分けの情報を指摘するのは無駄であろうし, 知らない場合でも一回指摘をすれば充分で, 同じ指摘を繰り返すことにはあまり意味がないと考えられる.

形態素解析における失敗部分の指摘 誤りの検出技術としては, 形態素解析を用いて解析辞書とのマッチングに失敗した部分（未登録語）, あるいは未登録語でなくとも不自然な解析部分を調べ, 形態素解析結果に対して特定の条件を満たす単語列を抜き出す仕組みが挙げられる [2].

タイプミス等のランダムな誤りや文字単位での漢字入力による誤りにはこのような方法で多くが検出できるが, かなりの過剰な指摘を含むことも事実である. こうした検出手段は実際には, システムのもつ辞書の不備や対象外の文法的現象（古語の活用等）を指摘することになってしまうことが多い. これを避けるためには事前の辞書の整備や活用語尾の接続情報の調整等が必要となってしまう.

発音によってソートされた単語リストの提示 ユーザ自身の誤りのチェック作業を支援するツールとしてはカタカナ語や漢字語を抽出して発音順にソートした語彙リストを表示するものがある [1].

特に語彙の変動の大きいカタカナ語に対しては辞書によるチェックが難しく, この種のツールは有効である. しかし思い込みによって文中の単語が全て誤っていた場合や, 一回しか現われない単語に対しては無効であることやチェックの対象とする単語リストの一つ一つを自分でチェックする必要があるという点で作業効率上の問題がある.

4 システムの設計目標

既にある程度述べたことであるが、システムの設計に際しての目標を下に示す.

- 様々な種類の表記誤りに対して信頼性の高い誤り指摘を提示する枠組の実現

  現在の状況ではユーザは誤りの種類に応じて様々なツールを選択し, 指摘の信頼性についても自ら判断しなければならない. 我々は再現率や適合率においてある程度質の揃った情報を様々な種類の表記誤りに対して提供できるような統一した枠組を目指す.

- データ整備／調整の容易さ

  対象とするテキストや入力手段, 校正基準への調整が容易に行なえるためにはまず項目毎の独立した調整ができるようなシステムの枠組が必要となるが, それだけでは不十分であり, データ整備の工数を軽減することも重要な問題となってくる. 我々は単語辞書, コーパス等の既存のデータベースからシステムで必要なデータを機械的な処理によって整備することを目標とする.

上述の設計目標を達成するために我々は校正支援システムを仮説の生成部と検証部の２つの要素を独立し

た要素とし, 互いに独立して動作する構成とした（下図参照）.

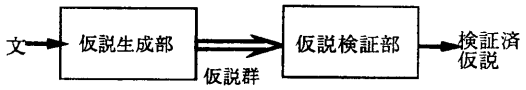

図 1: 仮説の生成と検証の概念図

仮説生成部は誤りの種類に関する知識を使って入力テキスト中の誤り部分と誤り内容についての仮説を生成し, 仮説検証部は語彙情報との照合によって生成された各々の仮説を検証し, その中から最適な仮説を選択する. これによってまず誤り種類に関する情報と語彙情報を分離して, 相互に独立した調整ができるようにした. 又誤り仮説は全て語彙による検証を得た後で提示されるために適合率の高い指摘が可能である.

仮説の生成と検証の概略としては, 先に述べた文字レベル, 単語レベルの各々に対して既存のツールの機能を一部組み合わせて, 次のような戦略をとることとした.

- 文字レベルの誤り

  形態素解析に失敗している疑いのある部分を抽出して, 文字としての混同に対する展開規則での仮説を生成する. それに対する検証は通常の単語辞書とのマッチングによって行なう

- 単語レベルでの誤り

  単語間の共起情報（複合名詞, 複合動詞）による検証を行なう. この場合は名詞の連続部分等複合語の候補のありそうなところを対象として単語単位の展開処理を行なう.

5 システムの基本的な構成とラティス構造の説明

我々の構築した校正支援システムの構成, 仮説生成と検証の枠組について説明する.

前節で述べた誤り仮説生成部は更に誤り部分検出部, 誤用候補展開部の 2 つから構成されている. 誤り部分検出部は入力テキストを形態素解析し, 誤りの存在する可能性のある部分を抽出する. 誤用候補展開部は対象とする部分テキストに対して展開規則を参照して誤りの内容についての推定を行ない, その推定結果をラティス構造と呼ばれる 2 端子グラフの形で提示する. テキストに対してラティスを対応させる処理を誤用候補展開処理と呼ぶ.

仮説検証部（正解語探索部）は単語辞書や正解語辞書（固有名詞, 複合語情報）との照合によってラティス構造中から最適（コスト最小）の経路を探索する. ラティス構造の例を下図に示す.

ここでは「安全保証の」というテキストに対して「保証」という単語が同音異義語誤りを含む可能性を

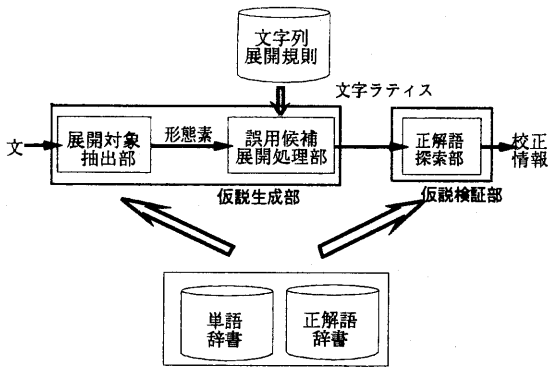

図 2: Joyner の構成図

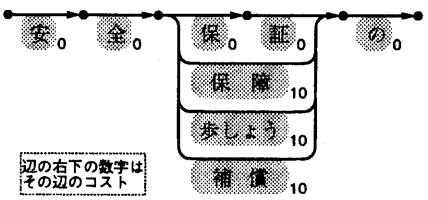

図 3: ラティスの例 (入力:「安全保証の」)

考えた結果,「保証」に対して「保障」,「補償」,「歩しょう」が綴の新たな候補として原文に付加されている.

ラティスの個々の辺は綴（文字列）と展開コストを持っている. このコストは展開処理の際にシステムが設定するコストであり, 各々の仮説の信頼性（又は誤りの確率）に関連する情報を伝える. .基本的にはコストの高低の順序関係は誤りの確率に対応し, コストが高いほど誤り確率が低くなる訳である. ここでラティス中の経路のもつコスト C の計算式を示す.

$$C = \Sigma(C_e + C_m) + \Sigma C_c$$

各々の経路のコストはラティスの各辺に付加された展開コスト C_e, 形態素の品詞に対応した形態素コスト C_m, 各形態素間の接続部における接続コスト C_c, の 3 者の合計から計算される. 例に示したラティス構造に対しての正解語探索を考える（辞書中に複合語「安全保障」が登録されていると仮定する）と下に示すような様々な経路が考えられる.

経路	C_c	C_m	C_e	合計
安全 | 保証 | の	20+20	0	0	50
安全保障 | の	20	0	10	40
安全 | 補償 | の	20+20	0	10	60
安全 | 歩しょう | の	20+20	0	10	60

この内,「安全保障」を含んだ経路が接続コストが少なくなることで最小コストとなり, 選択される. 選択

された経路が入力テキストの表記と違うために元のテキストに誤りがあったとみなし,対応する校正情報が出力される.

6 システム各部の詳細

ここでは前節の各構成要素の内部を更に詳説する.

6.1 展開対象領域抽出部

誤り部分抽出部では,入力テキストを一旦形態素解析して,その解析結果に対して誤り部分の抽出を行なう.正解語探索の対象とする複合語句は単語辞書内に登録しておくことを前提としている.これらの単語は単独の単語と認識され,(不必要な)単語レベルの誤りに対する処理は起動されない.

展開対象領域検出部

先に述べたように単語レベルの誤り候補の検出と文字レベルの誤り候補の検出を行なう.

単語レベルの誤り候補の抽出 ここでは名詞の連続部,動詞の連続部を複合品詞の存在可能な(すなわち単語レベルの誤りが存在し得る)領域として抽出する.

ex. 安全 保証 への 影響 を 差し 示す. (下線部が抽出対象)

文字レベルの誤り候補の抽出 ここでは正しい単位での切り出しのためにまず確実な境界によって区切られた区画への分割を行ない,その後,区画毎に単語の認識誤りをチェックして,認識に失敗している部分(誤り区画)を抽出する.

1. 確実な単語境界による分割

   活用語句の認識に失敗した場合,語幹部分が分離されて後続の平仮名部分全体に影響が及ぶことが多い(下例を参照のこと).このような境界を避けるために,単語境界中から確実な境界の基準として次のものを選択することとした.

   - 助詞「を」,「,」等の記号の前後
   - 平仮名から漢字に字種が変わる箇所

   ex. /彼(名)は(助)/黙(動)って(尾)うなづ(未知語)い(動)た(尾)/ (/が確実な境界)

2. 誤り区画のチェック

   一般に文字レベルでの誤りがある部分が必ず未登録語となる訳ではない.ここでは形態素解析の失敗の判断を未登録語の存在以外にも領域が細分化されたことで判断しており、このため次のような誤り区画の判断基準を導入している.

   - 未登録語を含む
   - 形態素解析結果が字種毎に設定した想定単語長よりも細かく分割された.

     (ex. カタカナ領域の想定単語長:4 漢字領域の想定単語長: 2)

ex. /テイル ランプ/を/点灯する/ (下線部が誤り区画)

6.2 誤用候補展開部

検出された複合語領域,誤り区画の各々に対して正解語(綴り)を含む解候補の生成を行なう.

誤り区画の展開

誤り区画に対しては外部の規則規則を利用して文字レベルのマッチングに基づく展開を行なう.展開規則は現在,漢字,カタカナ,平仮名の各字種に対応した個別の規則を持っており,対象字種で構成された部分に対しての展開を行なう.各展開規則は展開要素個々に対して展開コストを記述できるようにしており,誤り頻度に応じた柔軟なコスト設定が可能である.

カタカナ文字列に対する展開例を下に示す.

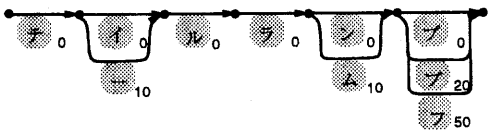

図 4: カタカナ部分の展開例

複合語領域の展開

複合語領域に対しては基本的には形態素辞書中に記述した単語展開データによる展開処理を行なう.

新聞校正に関する統計でも誤りのほとんどは単語対単語のものであることから,我々は誤用候補の対象を1対1の単語レベルに限り,情報は形態素解析辞書中の単語自体に予め付加しておくこととした.これによって展開処理は形態素結果に含まれる情報のみで行なうことができ,処理の負荷が軽減される.

単語綴	* 誤用候補情報
保証	保障,補償
差す	指す

表 1: 辞書記述例 (* が付加部分)

展開コストの設定に関しては,展開データ個々の設定はしておらず,単語展開データ全体に対する一括的なコスト設定のみとしている.

6.3 正解語探索部

我々は形態素解析の枠組として既に利用中であるCYK法を用いた探索法を拡張して,正解語探索の枠組に利用した.これによって形態素解析の探索に自然な形で正解語探索を融合することが可能となり,形態素解析におけるコスト等のパラメータを共有することができた.

形態素解析からの拡張は辞書引きをする表記の組み立てに関する制約をCYK表に表す方法と,その結果を用いて辞書引きを行う方法(CYK表の初期化)について行なった.その各々について述べる. CYK法による形態素解析の説明は省略する.

ラティス構造のCYK表への割り付け

文字列の形態素解析の場合, 文字列とCYK表との対応付けは文字列を先頭から一字ずつ割り付ける.OCR等で同一位置に対して複数の文字が対応する構造(matrix)が対象とされることがあるが,この場合は複数の文字候補を同一位置に割り付けるだけで対処できる.これに対してラティス構造では複数の文字列同士が置き換え可能な文字列として対応し得ることとなる.この場合に対応するために我々は元の文字列の何個の文字に対応するかを表す『仮想長』のデータを導入した.

図3に示したラティスでこの割り付けを行った様子を図5に示す. ここで元々の文字列に対応する文字は一文字毎に独立したエントリとして仮想長1と共に割り付けられているのに対して,展開の結果生じた文字列「保障」「歩しょう」等は展開結果全体を一エントリとして元の文字列の「保証」の文字長2と共に登録されている.展開結果全体が一エントリとなっていることは展開によって生成された文字列の内部に単語の切れ目がくることはあり得ないことを意味する.

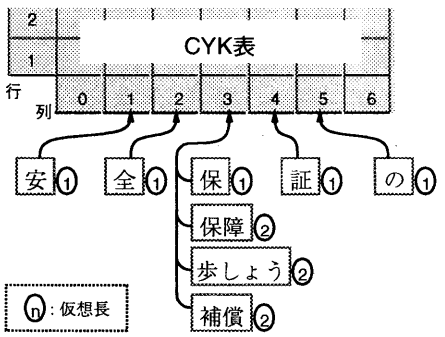

図 5: CYK表への辺の割り付け

辞書引き

従来の文字列の形態素解析の場合, CYK表の隣あった列(文字に対応)を組み合わせた文字列を表記として辞書引きを行っているが,上記のラティス構造に対する割り付けを行なったCYK表の場合,辞書引きの対象はもう少し複雑となる.

すなわち表記を合成する際にある列iに後続させ得るて使って良い列jは, 列iから列iの仮想長だけ進めた列となる.例えば上の例で列3に登録された「歩しょう」を辞書引きの対象文字列の一部として選択すると,仮想長2が指定されているために次は列5の部分に割り当てられた「の」がその後に続き得ることとなる.

図5で列2に割り付けた「全」から始まる辞書引きの例を表1に示す.

表 2: 辞書引きをする表記

表記: 辺の表記(位置, 仮想長),···	結果を登録する行
全(2,1)	1
全(2,1), 保(3,1)	2
全(2,1), 保(3,1), 証(4,1)	3
全(2,1), 保(3,1), 証(4,1), の(5,1)	4
全(2,1), 保障(3,2)	3
全(2,1), 保障(3,2), の(5,1)	4
全(2,1), 歩しょう(3,2)	3
⋮	

7 データの整備作業について

先に述べたように我々は調整の容易さの実現のためには整備工数の軽減が不可欠だと考えており、このためシステムの各種データの整備を既存のデータベースの情報を元にして機械的な操作によって行なう枠組を作り上げた.

以下では枠組を利用して特定の目的のために行なった整備作業について報告する.ここでは日本語入力手段としては一般的な仮名漢字変換ソフト全般に対応できることを目指し、テキスト分野としては(校正済みの大量のテキストの入手を考え)新聞の一般記事を選択した.

7.1 展開データの整備

文字展開規則の整備

- 漢字表記展開規則

  制約条件が少なく,発生する仮説数が一般に多い文字レベルの誤りの中でも特に漢字表記は展開文字の数が一般に多いためにためにメモリスペースに対する負荷が大きい. このため,発生させる仮説は新旧字体の対応のみに抑えた.

ex. 浜 ⟶ 濱

- 平仮名表記展開規則

  送りがなの揺れ, 漢字と平仮名表記の間の表記揺れは誤りパターンがかなり限られるために, 基本的に校正辞書への誤り語の登録によって扱うこととし, 展開規則では二重母音の表記, 「づ」と「ず」等の表記の揺れを扱うこととした.

  ex. こう ⟶ こお

- カタカナ表記展開規則カタカナの表記誤りについては基本的には表記の揺れに基づくものを対象とし, (半) 濁音の付け忘れ等の簡単な認知誤りを更に追加することとした. 単語辞書中のカタカナ語彙, 新聞社のハンドブックの記述を参考として約 70 の規則を抽出した.

  ex. エイ ⟶ エー

単語展開データの整備

基本的には同音異義語誤りを対象とした整備のみを行なった. 仮名漢字変換においては同音の別単語への変換誤りが全て同一の確率で発生し得ると仮定した. データの整備においては形態素解析用の基本単語辞書 (読み付き) を対象とした. データ整備の手順について説明する.

1. 単語辞書中の自立語を品詞によって名詞類, 各活用タイプ毎の用言類等に荒く分類.
2. 各カテゴリの内部での同音単語をグループ化して, 相互に展開可能な同音異義語データを生成.
3. 複数の読みを持つ単語については各々の読みによる同音語展開データをマージ.
4. Joyner 用単語辞書の各エントリに対して表記と品詞によって対応する同音語データを検索し, 検索されたデータを各単語エントリの属性として書き込んだ.

7.2 検証用データの整備

ここでは正解語探索時に参照される辞書データの構成について説明する. 基本的な語彙に対して, 人手で整備した組織体名データ, コーパスから抽出した複合語データ, カタカナ語データ (正解語) を加えることによって検証能力の向上を図っている. 語彙構成を次に示す.

- 単語辞書 (115,600 語)

  形態素解析用の基本辞書で固有名詞も含む.

- 組織名データ (23,000 語)

  変動の激しい組織名のデータを人手で整備したものである. 日本国内の一部上場企業, 公共機関等の正式名称を登録した.

- 正解語データ

  複合名詞, カタカナ語彙を毎日新聞の記事コーパスから抽出した. 以下の節ではその整備作業について述べる.

正解語データの整備

複合名詞等の単語の共起情報は特に漢語の同音異義語誤りの訂正に重要な役割をもつ. 又, 変動の激しいカタカナ語彙情報もカタカナ表記の揺れ訂正のために不可欠な情報である. 我々は毎日新聞の 2 年分の記事データを利用し以下の手順によってこれらのデータの抽出を行なった.

1. 記事による選択

   人事関係の記事等抽出される人名が部課長のクラスで対象人員が多く, 他の人名との混同の可能性がある. このために人事異動関連の記事は正解語の抽出対象から除いた.

2. 見出し部分の除外

   記事の見出し部分は助詞の省略によって名詞句の切れ目の認定が困難, 「ア社」等の省略名の多用等の理由で正解語としての利用度は低いと判断し抽出対象から除外した.

3. 複合名詞, カタカナ語の抽出

   複合名詞の内部に含まれ得る品詞の連続 (名詞, 接頭語, 接尾語等), カタカナ未登録語部分を抽出する.

4. 複合語に対するスクリーニング

   複合名詞に対して数量表現や肩書無しの人名表記等の正解語としての利用度が低いものを除く措置をした. 又, 複合語については各構成単語について同音の別単語が存在するかを調べ, 構成単語が同音語を持たないものについては除いた.

5. 頻度による絞り込み

   同一表記の単語の生起回数をカウントして生起回数による絞り込みを行なった.

8 評価

我々の目標とした誤り検出能力の向上を再現率と適合率の点から評価した結果について述べる.

8.1 複合語中の同音異義語誤りの訂正能力

まず, 複合語中の同音異義語誤りの訂正能力に関する評価を行なった.

正解語データとして我々は毎日新聞の 1991,1992 年分の記事データを利用し前述の手順によって語彙数約 90,000 のものを用意した. 更に時事用語として 150 語の複合語を新聞記事からランダムに抽出し, 更にそれに対する変換誤りを人為的に生成してこれらをテストデータとした.

これらに関しては次のような結果を得た.

	件数	処理例
訂正成功	93 件 (62%)	新王子生死 → 新王子製紙
無訂正	54 件 (36%)	生物主
誤訂正	3 件 (2%)	管掌券 → 管掌研

表 3: 同音異義語誤りの訂正結果

無訂正の場合はそのほとんどが訂正の基準となる複合語が辞書中になかったことによるものであるが, 漢字未登録語の展開データの不足によるものが 10% 程あった. これらは単語として登録されていないような単漢字への展開データである.

これとは別に人手で抽出した時事用語に対して, 正解語辞書の語彙数を増やした場合にカバー率がどの程度上昇するかを見たものを示す.

語彙数 (万)	15	17	19	21	23
カバー率 (%)	63.9	69.4	72.6	73.9	76.5

表 4: 正解語の語彙数とカバー率

8.2 カタカナ語句の表記のゆれの訂正能力

カタカナ語句については社内の新聞用校正データ (誤り表記と正表記のペア) を対象としてカタカナ表記の揺れに関するデータを約 3400 件抽出して, それらをテストデータとして校正元データから校正先データへの展開ができたかを調べてみた. その展開の成功率は約 90% で約 3000 件のデータに対しての展開に成功した. 失敗した部分は母音に関する揺れが主であった.

ex. レファランス ⟶ リファレンス

8.3 過剰訂正の割合

対象テキストとして日経新聞記事 222 文 (11274 文字 (全て 2byte 文字)) を選択し, 校正システムに入力した. 基本的には誤りはない筈のこれらの記事に対しての指摘された誤りの数は 14 件であった.

過剰指摘の原因についてはほとんど (12 件) が辞書エントリの不足によるものであった. 下に例を示す.

大証 ⟶ 大小

このような短い単語が未登録の場合は同一の読みを持つ他の単語に置き換えられる可能性が高い. こういった漢字語句もコーパスからの抽出のための仕組みを考えていく必要があろう.

9 まとめ

現在このシステムは既に富士通のワープロソフト OASYS/Win95 に組み込まれ, まず一般的なワープロ・ユーザに向けた製品として利用されている. 校正支援システム単体での性能評価ではマシンは FMV590T2 (Pentium), メモリ 16M,OS として Windows-NT3.5 を利用した処理で初期化に 0.4 秒, 総処理時間として 61.8 秒かかっている. 形態素解析のみに処理に比べて数倍の処理時間でこのような処理が実現できた.

システムの誤り指摘情報の信頼性、分野移行の容易性はある程度証明できたと我々は考えている. 今後はデータ整備によるシステムの検証, 述語を中心とした共起関係の導入による訂正範囲の拡張を計画している.

- データ整備

  大量の正解語データによるシステムの限界性能の評価, 各種の FEP に特化した展開データによる展開の精密化を行なう又, 特に平仮名部分規則の拡充によりでのランダムなタイプミスの検出能力を拡大させる.

- 検証能力の拡大

  現在は単語レベルでの誤りの検証は複合語語彙との照合によって行なわれているために単独で出現する名詞類は枠組から外れてしまう. これに対して述語と格要素間の共起関係をチェックする仕組みを導入することによって対処を目指す

最後に, 新聞の校正過程や原稿の誤り例についての詳細なデータや助言を頂いた福岡克氏, 原稿のチェックや技術的な問題点について指摘して下さった西野氏に感謝する.

参考文献

[1] 橋本他: “誤った日本語に気づき始めたワープロ第 3 部 文書校正支援”, 日経バイト 1995 年 1 月号

[2] 脇田早紀子他: “日本語校正支援システム FleCS-ミスタイプ検出について”, 自然言語処理,97-19,pp.135-142(1993)