新聞社向け校正支援システムにおける文書校正の新方式

New Document Proofreading Method in Proofreading Support System for Journalist

あらまし

富士通では，株式会社富士通研究所と共同で，新聞社の記事制作業務を支援するシステムとして，新しい方式を採用した校正支援システムを開発した。

従来の校正支援システムでは，校正実例や経験的に誤りやすいとされる表記のデータベースを手作業で構築し，そのデータベースを参照することで誤りの検出を行っていた。この方法では，誤りやすいとされている表記が誤って出現した場合にしかある程度の効果を上げることができない。また，長い時間をかけて開発しても，一般的に出現する誤りを十分に検出できるだけの情報を蓄積することができない。

これらの問題を解決するために新しい校正手法を開発した。さらに近年利用しやすくなった大量の電子化されたテキストデータから校正処理に役立つ情報を抽出することで校正精度の向上を目指した。

Abstract

Fujitsu Limited and Fujitsu Laboratories Limited have jointly developed a new type of proofreading support system that supports article creation tasks in newspaper companies.

In conventional proofreading support systems, a database of proofreading examples and incorrect phrases was constructed manually and errors were detected by referencing the database. However, this method can produce useful results only for specific mistakes. Also, even if much time is spent developing the system, it is not possible to accumulate sufficient information to achieve a satisfactory level of detection for common errors.

To solve these problems, a new proofreading technique was developed to improve the proofreading accuracy by extracting information that is useful for proofreading processing from the large amount of electronic text data that has recently become accessible.

足立　顕（あだち　あきら）

1992年東海大学理学部情報数理学科卒。同年富士通入社。以来新聞CTSパッケージの開発に従事。
情報出版システム統括部第一システム部

新聞社向け校正支援システムにおける文書校正の新方式

ま え が き

ほとんどの新聞社では，富士通のPRESS[1]（PRogressive Editing Support System）に代表される，組版システムであるCTS[2]（Computerized Typesetting System）を導入することで，新聞制作業務のほとんどがOA化されている。記事制作についても，投稿記事など一部の例外を除けばほぼ100%の原稿がワープロを使用して作成されている。

ワープロで記事を作成する場合に使用するのが「かな漢字変換機能」である。最近では，かな漢字変換精度も高く，入力の方法次第では，ほとんど誤りのない変換結果を得ることができるまで進歩している。

しかし，入力者の「くせ」により，かな漢字変換機能の能力が十分に生かされていない場合が多い。そのため，かな漢字変換ミスの代表ともいえる，同音語の選択ミス，言い換えれば同音語誤りが発生する。

また，かな漢字変換機能で使用する辞書は，新聞表記基準に合わせ専用に構築されたものではない。そのため，新聞記事の表記・表現基準に合わない表記に変換されてしまう可能性がある。

新聞社では，記事の内容もさることながら，日本語の質も新聞の品質として評価されるため，一般的な表記基準より厳しい規定を定めている。さらに，通信社や友好社から配信された記事を，自社の記事と同じ紙面に掲載した場合に表記の不統一が発生しないよう，表記基準を忠実に守る努力をしている。ところが，画面上の文字は，見た目には整っており読みやすい。この読みやすさが原因して，誤りが見つけにくいという副作用が生じている。新聞社では，新聞としての品質を維持するために記事中に隠れている誤りを見つけ出すために多くの時間と労力を費やしている。

これらの背景から校正を支援するシステムに対する期待は大きい。

その要望を受け，富士通は1996年に翻訳用の日本語処理機能の一部を利用した校正支援システム[3]を提供した。

しかし，翻訳用の日本語処理機能は誤りが含まれている文書を解析することを目的としていないため，「かな漢字変換ミス」のように特徴がつかみにくい誤りを十分に検出することができない。そこで，新しい校正処理の枠組みを開発した。

また，近年利用しやすくなった電子化された大量のテキストデータから校正に役立つ情報を統計的に抽出することで，誤りの検出精度の向上を目指した。

本稿では，機械的に取り出した情報を利用した新しい校正手法と校正支援システムとしての今後の課題について述べる。

従来の校正手法

日本語は欧米などの言語と異なり単語の区切りが不明確である。単語レベルの処理が必要な場合，入力文そのものを直接処理することができない。そこで，入力文を単語（形態素）単位に分割する処理（形態素解析）を行う必要がある。

従来の校正手法は，形態素解析後に取得される形態素に付加された校正情報（形態素が誤り表記であることを識別するための情報）を検索することで誤りの検出を行う方式である。

また，形態素と形態素の係り受け関係を解析（構文解析）し，係り受け関係上の誤りパターンの校正情報を検索することで誤りを検出しようとする試みも行われている。

1996年に製品化した校正支援システムは，この形態素解析と構文解析を採用したシステムであった。

従来の校正手法の問題点

従来の校正手法は，誤りであるという情報，すなわち校正情報により誤りを検出していることは述べた。誤りであるという情報は，人間が経験的に判断し決定しなければ作成することができない。言い換えれば，校正情報の作成は，人間の経験に基づき手作業で作成しなければならないということになる。

校正情報に依存するこれらの方式では，誤りが有限であり，かつ誤りであることが特定しやすい場合に限り効果的な方式であるといえる。しかし，誤りであることが特定しにくい，すなわち形態素と形態素の組合せ関係上で誤りとなり，さらにその組合せが多い場合では，大量の校正情報が必要となるために，事実上，誤りの検出に対しほとんど効果を上げることができない。とくに，校正情報で対処できない代表が次章で述べる同音語の変換ミスである。

同音語変換ミス

最近のかな漢字変換機能は，「文」や「複文節」で入力すると同音語の使い分けを高い精度で漢字に変換することができる。しかし，入力者自身の「くせ」により，その入力方法は様々である。「単文節」や「単語」単位で変換する人が多く，かな漢字変換機能を十分に活用できていないといえる。このような入力方法で変換処理を行うと図-1のような結果が得られる。

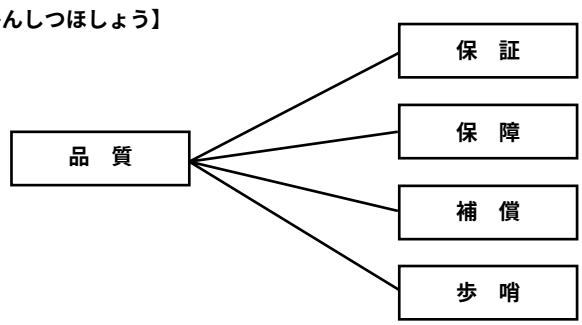


図-1　単語で変換した場合の表記の組合せ例
Fig.1-Examples of translation from Kana character to Kanji character.

変換結果が，どの組合せになるかは，かな漢字変換機能の種類によって様々な傾向があり，特徴づけることはできない。さらに，同じかな漢字変換機能であっても学習機能などが機能している場合，以前に処理・選択した状態により変化する。

また，変換精度とは別の問題として入力者の状態によっても状況が変化する。例えば，思考しながら入力している場合である。一般に思考しながらの入力では，書いたり，消したりという編集操作が頻繁に発生する。そのため思わぬ場所に誤りが紛れ込むことがある。このような状況では，単に，かな漢字変換の変換能力が向上しても，変換ミスを解消することは困難であるといえる。

このようにして生じた誤りは，文中のほかの表記との組合せ(意味的なつながり)により誤りとされる。

同音語変換ミス検出の新方式

これまでの説明により，同音語変換ミスを校正情報で十分に検出することができないことはいうまでもない。

以降では，富士通研究所と共同で開発した新しい取組み[(4)]について説明する。

● 同音語変換ミスの可能性の検出

同音語の変換ミス(同音語誤り)が含まれている部分を形態素解析すると，ほかに同音語を持つ形態素(例えば同音異義語が存在する表記など)が単独で検出される。隣接する形態素の種類によっては，本来分割されずに，一つの形態素として解析されるべきものである場合がある。例えば，固有名詞や熟語など，広い意味で捕らえた場合の複合語である(図-2)。

このような場合，分割されて解析されるのは不自然であり，むしろ誤りが含まれているために分割されてしまったのだと考える方が素直である。

● 正解語探索方式

同音語変換ミスの可能性の検出については前述したが，あくまで可能性であり，すべてを誤りと判定するのは強引である。なぜならば，形態素解析を行うための辞書に未登録であるために分割された場合も同様の解析結果が得られるためである。

そこで，図-3に示す方式を取り入れることで，誤りであることの確度を高めた。

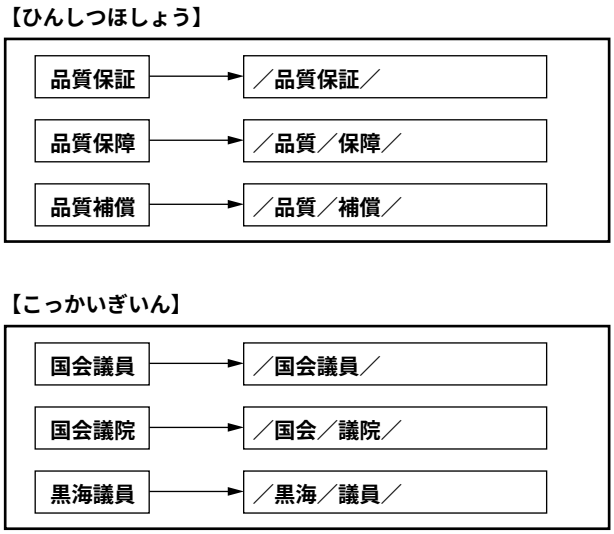
この方式は，同音語変換ミスの可能性のある部分に，別の同音語を持つ表記が含まれていた場合，同音語部分を置き換えて再度辞書を検索するというものである。同音語を置き換えて辞書中の形態素と一致した場合，置き換える前の形態素より後の形態素の方がより正解に近いと判断できる。したがって，正解語として検出された形態素が，誤りに対する校正候補となる。

同音語を置き換えて辞書を検索する行為は，変換ミスを考慮しながら辞書中の正しい形態素を検索するという動作に似ていることから正解語探索方式，または，誤りの情報を蓄積した辞書(誤り語辞書)を使用した従来の方式の対語として正解語辞書方式と呼んでいる。

● 正解語の蓄積

この方式を採用することにより，誤り個所の検出と校正候補を取得することができ，同音語変換ミスを検出するために，従来の校正支援システムが使用していた校正情報を大量に作成する必要はなくなった。

しかし，新しい方式では，従来の校正情報に変わり，大量の正解語が必要となった。校正処理の対象となる文



矢印の左側：形態素解析前の状態
矢印の右側：形態素解析後の状態（単語の区切りを「／」で表記）

図-2　不自然な形態素解析結果
Fig.2-Result of analyzing Japanese sentence which has mistaken.

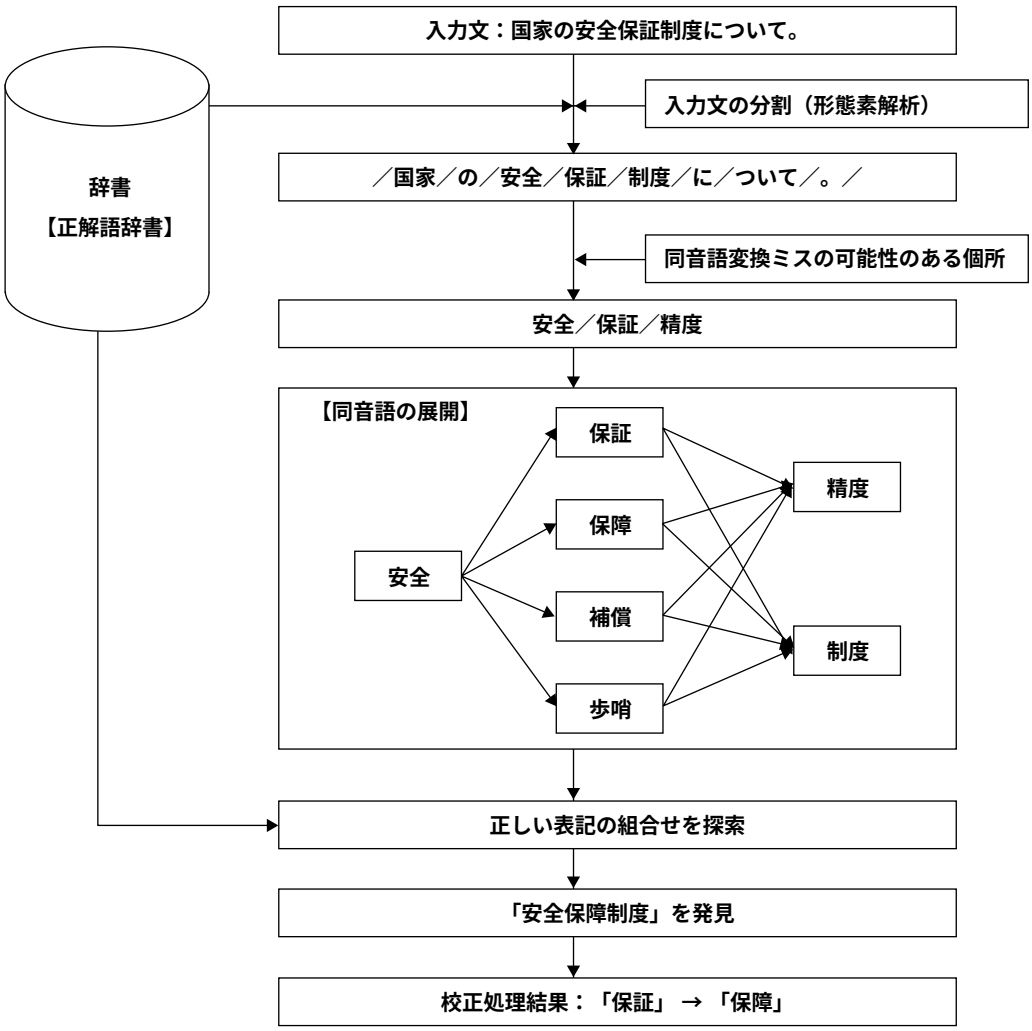


図-3　処理の流れ
Fig.3-Flow of transaction.

に対して十分な正解語が蓄積されていなければ，単なる未登録な形態素と同じ扱いになってしまう。

そこで，使用したのが電子化された大量のテキストデータである。校正情報を作成する場合と異なり，正しい表記の情報はテキストデータを解析すれば，機械的に大量に取り出すことが可能である。このようにして得られた正解語を必要な語数取り出して辞書にすることで，正解語の蓄積に対する問題を解消した。

● 正解語探索方式による効果

正解語探索方式を採用したことで，同音語変換ミスを幅広く検出できるようになったことは言うまでもない。特筆すべき点は，辞書情報の整備に関する効果である。

本稿で論じている校正手法は，新聞社の新聞制作業務を対象として開発した校正支援システムに搭載されている。全国的なニュースであれば一般的な時事用語を蓄積しておけばよい。しかし地方独自のニュースの場合，個別に表記を追加しなければ未登録の状態のままである。従来では，このような形態素を辞書に登録したり，校正実例などをもとに校正情報を作成したりしていた。そのため導入までの準備作業に多くの時間を必要とした。

一方，正解語辞書方式を採用した場合では，電子化されたテキストデータがあれば自動的に大量の情報を取り出すことができる。

近年新聞社では，新聞紙面としての記事の利用方法以外に商用データベースとして記事を活用する傾向があり，簡単に大量のデータを採取することができる。そのため，従来と比較して短時間で新聞社の条件にあった辞書を構築することが可能となった。

これにより，初期導入時の作業負荷を軽減させることが可能となり，各新聞社に適用しやすくなったという効果が得られた。このことは同時に最新の情報に辞書を更新することが容易になるということを意味している。

今後の課題

新しい校正手法について述べてきたが，現実世界に存在する誤りの一部をカバーしているにすぎない。以降は，校正支援システムの今後の課題について述べる。

● 正解語探索方式の拡張

現在実現している正解語探索の機能は，複合語(隣接した単語間の関係)の範囲である。同音語の誤りが発生する可能性は，複合語の範囲にとどまらず，構文・意味的なものまで幅広く存在する。

構文・意味的に誤っているような状態は，経験的に人間が間違いやすいとされている誤りで，その特徴が明確な場合を除くと一般的には誤りの特徴が捕らえにくい。複合語の場合と同様に表記間の組合せで誤りとなり，入力する状態によって異なるからである。

そのため，校正情報などを活用して検出する従来の方式では対応することが難しい。

しかし，本稿で論じた新しい校正手法を用いて構文，意味的な誤りを検出するためには，構文，意味的なつながりの情報を持つ正解語の辞書を構築する技術も必要である。そのため，単純に当てはめることができない。

今後は，正解語探索方式の特長を生かしながら処理の枠組みの拡大を図りつつ，同時に正解語収集の枠組みを開発する必要がある。

● 辞書の寿命

新聞に掲載される「ことば」や用語は，日々増加する傾向がある。今回の大量の電子化データを利用した取組みでは，従来の手作業により辞書を構築していた場合と比較にならないほど，幅広い表記を蓄積した辞書の実現を可能とする。そのため，ある程度の表現の変化には追従することができる。しかし，新たに出現した固有名詞などに対しては，効果がない。こまめに辞書に単語を蓄積すればよいのだが，その単語の寿命が短い場合があり，利用者側で積極的に辞書登録する動きは見られない。

半年，１年ごとに電子化データから一括して取り出すことで，最新の語彙を維持することが可能ではあるが，これだけでは突発的な状況に対応できない。

校正の立場から言えば，新しく出現した表記を人間的判断抜きにして正解語として蓄積することは，困難であるといわざるを得ない。しかし，「ことば」を扱うシステムとしては取り組んでいかなければならないテーマであると考えている。

辞書の寿命を延ばすための取組みとして，効率よく正解語を蓄積する枠組みと，データベースなどの大量の電子化データから一括で取り込む方式をうまく活用する研究を行い，ことばの変化に柔軟に対応できるようにする必要がある。

● 事実関係の誤りへの取組み

新聞社は，報道という立場で速報性と正確さを必要としている。その反面，紙面のページ数増加など校正をしなければならない範囲も広がっている。正確さという点で，事実関係の誤りは致命的であり，中でも固有名詞の取り扱いはもっとも神経質にならなければならない。

現状の校正手法では，複合語の一部である固有名詞中の同音語変換ミスを検出できるまで進歩した。

しかし，利用者である新聞記者が期待する事実関係の誤り検出とは，実世界との整合性の検証であり，多くの知識を必要とする。しかし，記事データベースなどを用いた知識習得に関する取組みが，情報処理学会などで始まったばかりである。すぐに実用化できるレベルに成長するとは考えにくいが，この技術が事実関係の誤りを校正する重要な基礎技術になるものであると考える。

この技術をいかにして校正に利用するか，研究する必要があると考えている。

む　す　び

現在実現されている校正手法は，人間により校正される誤りの一部を検出できるに過ぎない。人間が校正作業を行う場合，単に表記の誤りの修正だけでなく，文章の構成(読みやすさ)や事実関係に関連した内容の校正を含んでいる。

コンピュータに人間同様の能力を求めることは，コンピュータが誕生してからのテーマであるが，人間同等の校正作業ができるようになるまでは，まだまだ解決しなければならない問題が多い。さらなる調査と校正手法の研究が必要である。

参 考 文 献

(1) 山口，野池，遠藤：PRESS/FXシステム．*FUJITSU*，44，5，pp.432-439(1993)．
(2) 西山，榎本，山田：新聞社向け編集ワークステーションシステム．*FUJITSU*，45，3，pp.233-239(1994)．
(3) 足立，安永：翻訳システムATLASの校正支援システムへの適用．*FUJITSU*，47，4，pp.310-315(1996)．
(4) 松井ほか：日本語校正支援システム(Joyner)の研究について(1)．情報処理学会第52回全国大会，分冊(3)，2J-4，pp.283-284，1996.