自然言語処理 69－2
（1988. 12. 6）

文章校正支援機能における日本語解析

小山紀子，斎藤裕美，小林賢一郎[+]，小山和雄[++]
(株)東芝 情報通信システム技術研究所，
[+]東芝オーディオ・ビデオエンジニアリング(株)，[++]東芝コンピュータエンジニアリング(株)

本報告では、誤入力や基本的な日本語の誤りを検出することを目的の一つとして開発した、文章校正支援システムにおける日本語解析処理について述べる。

文章校正支援システムでは、日本語を解析するそれぞれの段階において、形態素解析を利用した単語単位，誤り指摘規則を利用した文単位，文章単位の誤り検出を行なう。

形態素解析による誤り指摘の結果を、論文の一次入力データを使って評価した。入力が正しいのに誤っていると指摘した部分の大半は、非文か辞書にない特殊な単語を使用している部分であり、形態素解析の不備のために誤って指摘した部分は１％に満たなかった。また、誤入力の指摘も、誤入力部分もしくはその充分近くを指摘できることがわかった。

JAPANESE ANALYSIS IN AUTOMATIC TEXT REVISION GUIDANCE SYSTEM

Noriko KOYAMA, Hiroyoshi SAITO, Ken'ichiro KOBAYASHI[+], Kazuo KOYAMA[++]
Information & Communication Systems Laboratory, TOSHIBA CORPORATION,
[+]TOSHIBA AUDIO VIDEO ENGINEERING Co., Ltd., [++]TOSHIBA COMPUTER ENGINEERING Co., Ltd.
70, Yanagi-cho, Saiwai-ku, Kawasaki, 210 JAPAN

We report Japanese analysis in the automatic text revision guidance system which is developed to detect errors in Japanese sentences.

It detects errors at each level of Japanese analysis, word level using morphological analysis, sentence level using error detecting rules, and text level.

We evaluated the result of error detection by morphological analysis using first input of papers. Most of all detection errors are due to incorrect Japanese or special terms. Detection errors which are due to inability of morphological analysis are less than 1 %. Our system can point out input errors themselves or positions which are near enough to them.

１．はじめに

日本語ワードプロセッサの普及により、文章作成が機械化されるようになった。以前のように紙に下書きし、紙の上で推敲して清書していた時代とは異なり、直接ワードプロセッサで書き始める場合も多い。また、ワードプロセッサなどの日本語入力機器は、キーボードやタブレットから入力するものがほとんどでタイプミスなどの誤入力は逃れられない。

このような初歩的な誤入力や基本的な日本語の誤りは、機械的に検出が可能なものが多い。人間では見逃しやすい誤字・脱字，送り仮名や基本的な構文の誤りなどを検出する機能は、今後のワードプロセッサなど文章処理機器にとって、必要不可欠なものになると思われる。

このような誤りを検出することを目的の一つとして、文章校正支援システムを開発した。日英翻訳システムのために開発された日本語処理技術を応用し、解析する過程と解析結果の両方から日本語文の誤りを検出する。本稿ではこの文章校正支援システムの概要と、一次入力後の誤入力の検出について、情報処理学会全国大会の論文を使用した評価実験の結果について報告する。

２．誤り検出の概要

検出すべき誤りとして、以下のようなものが考えられる[1)]。

(a) 誤字・脱字，誤入力などで単語として正しくないもの
(b) 送り仮名などが誤っているため、活用として正しくないもの
(c) 付属語や自立語など、単語の接続が正しくないもの
(d) １つの述語に主語と考えられるものが２つ以上あるなど、構文的に正しくないもの
(e) １文が長すぎたり、連用形が続いたりして読みにくいもの
(f) 同一文書中に、送り仮名や平仮名表記などの異なった表記が混在しているもの
(g) 同一文書中に、異なった文体が混在しているもの
(h) 括弧の対応がとれないなど、表記上誤っているもの
(i) システムやユーザが使用してはいけないと指定したもの

以上のものは現状技術で実現できるが、これ以外に以下のものがあり、これらは現在では限定して実現している。

(j) 意味的に正しくないもの
意味的な解析は、まだ局所的なレベルにとどまっている。完全な解析をすることは理解をすることになり、現状の技術では実現されていない。ここでは局所的な意味解析に限定し、用法などの誤りの一部を検出することにする。

(k) 敬語の用法が適切でないもの
敬語は日本語の中でも最も難しいものの一つであり、「世界」を持っていないと正しい使い方はできない。現在では、主語が一人称や二人称に限定した場合に多少できないこともないが、不完全なものであり、実用的でない。

3．誤り検出機能

文章校正支援システムは、以下の段階でそれぞれ誤り検出を行なう[2)]。

(1) 単語単位で行なう形態素解析
(2) 文単位で誤りを指摘する誤り指摘規則の適用
(3) 文章単位での表記上の誤りを指摘する方法

各々の段階で、それぞれどのような誤りを検出するかを前節で述べた検出すべき誤りの項目と関連して説明する。

(1) 形態素解析（単語単位）

形態素解析は、入力された日本語文を単語単位に分割する機能を持つ部分である。日本語文は英文などとは異なり、単語と単語がスペースで区切られていないため、単語に分割することが重要な役割を持つ。

単語の切れ目が明白でないため、文によっては単語分割の解釈にあいまいさが生じる場合もある。これらのあいまいさをできるだけ少なくしたり、一意ではあっても誤った解釈を棄却したりするために、さまざまな方法を用いている。この際に適用されるいろいろな規則や処理は、それ自体入力日本語文のチェックになっている。入力された日本語文が誤っている場合は正しい解釈ができなくなるということであり、それを誤り検出に利用する。

① 辞書検索部により語彙辞書を検索し、語彙辞書に未登録の単語を文章中から検出して指摘する。（(a) を検出）

辞書検索により品詞，接続情報，属性などを検出する。この情報を利用して文章を単語単位に分割するが、検索できなかった単語を未登録語（語彙辞書からでは単語として認められない語）として処理し、誤りとして指摘する。

② 活用辞書を参照することにより、活用語尾の解析を行ない、活用形を決定し送り仮名の誤りを指摘する。（(b) を検出）

検索された各単語は、活用辞書を参照して活用語尾の解析を行なう。動詞、形容詞などの用言は、このときに活用語尾を含めて単語として扱う。接続情報は活用形によっては変更される場合もある。この時点で送り仮名や活用の誤っているものは無効にし、有効な単語がない場合はその部分を誤りとして指摘する。

③ 各単語の接続情報を用いて単語間の接続を検定する。文節の最後の単語に付いては終了条件も考慮にいれて、単語間の接続誤りの指摘を行なう。（(c) を検出）

辞書検索により得られた接続情報により各々の単語間の接続を検定する。検出された個々の単語は各々いくつかの系列を持ちながら自立語と複数の付属語の接続により文節を形成し、文節の接続が文を形成している。ここで各々の単語間の接続条件が満たされない場合には、その単語列を無効にし、有効な単語列がない場合は、その単語を誤りとして指摘する。

④ あらかじめ語彙辞書に登録してある、使用してはいけない語や誤った単語が使用されている場合に、その語を指摘する。（(i) を検出）

語彙辞書にはあらかじめ、使用してはいけない語や使い方を誤った語など、単語分割でその語が使われること自体が誤りである語が登録されている。正しく形態素解析が行なわれていても、その解析結果にそれらの語が使われていれば、その部分を誤りとして指摘する。

(2) 誤り指摘規則の適用（文単位）

文節や単語の中での誤りを指摘することは形態素解析により可能であるが、単語間の接続判定は基本的に二項関係であることもあり、形態素解析では正しいと判断される文節と文節の関係に関しては誤りを指摘することができない。そこで、形態素解析によって得られた文節に対し、文節間の関係の誤りを記述した誤り指摘規則を参照して、誤りを検出する。

誤り指摘規則は、一般的な規則と語彙に依存した規則とから成る。一般規則は形態素解析により得られた文節系列の各文節に適用される。文節情報に語彙規則がついている場合には、同様に語彙規則をその文節に適用する。

誤り指摘規則は、主に構文的な誤り（(d)）を検出するために用いられる（但し、構文的な解析を完全に行なっているわけではないので、簡易版となる）。構文的な誤りを検出できる表層的な情報を抽出し、それを規則として書きくだすため、ヒューリスティックな性質を持つ。

語彙規則は、その単語に特有な規則を記述したもので、構文的な誤りだけでなく、局所的な意味の誤り（(j)の一部）を検出する規則も記述することができる。

文の長さなどで読みにくさを測る機能（(e)）も各文単位で行なう。

(3) 表記上の誤り指摘（文章単位）

文章全体を考えた場合に、括弧の対応がとれていないときにそれを指摘する（(h)）ことは有益である。特に長い範囲を括弧でくくってある場合、人間が見ていたのではその対応を見落としやすい。入れ子になっている場合に、その順序や深さをチェックすることも有効であると思われる。

(f)，(g)も、文章全体としてみた場合に指摘すべき誤りである。

文章校正支援システムの日本語解析部の構成を図１に示す。

４．評価実験の概要

この実験は、形態素解析による誤り指摘の結果を評価するために行なった。

実験に使用したのは、情報処理学会第３２回全国大会の論文集から３０８の論文を選び、かな漢字変換で入力できる日本語ワードプロセッサで入力した、一次入力のデータである。入力時にオペレータが気がついた部分は訂正してあるが、それ以外の部分については目視チェックおよび訂正はしていない。

形態素解析によって指摘された部分（但し、使用してはいけない語としてあらかじめ辞書に登録してある部分を除く）について、その原因別に分類して評価する。したがって、実際は誤入力であるのに指摘されなかった部分についての評価は行なっていない。

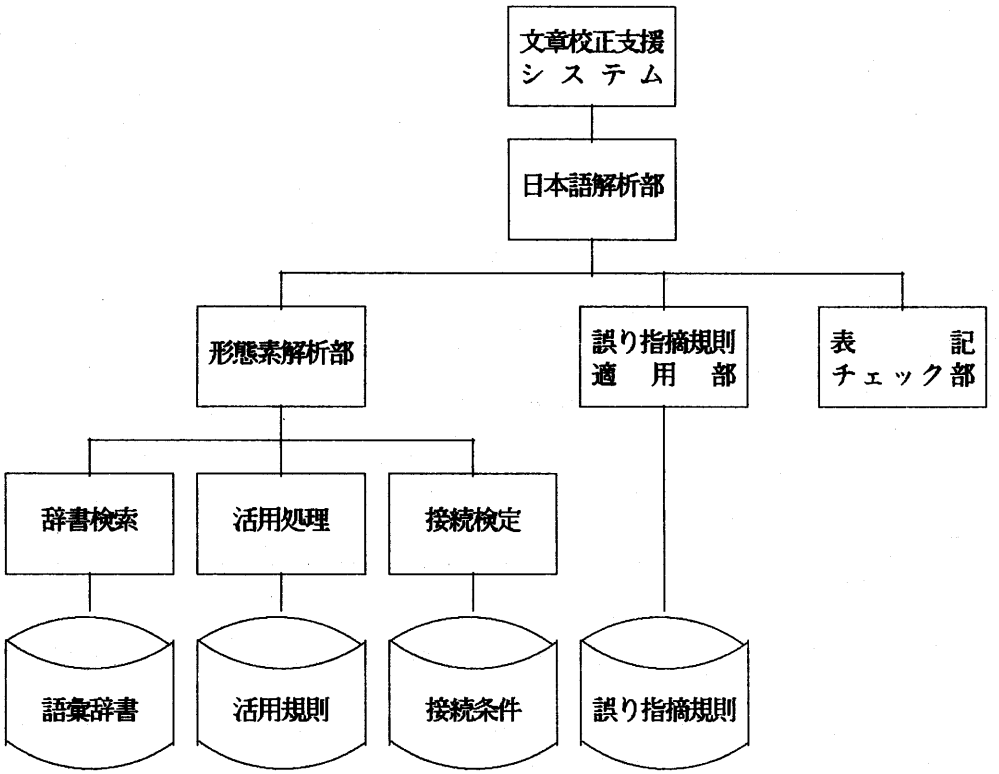

図1　文章校正支援システムの日本語解析部の構成

5．評価実験の結果

３０８論文から誤入力を指摘した部分のうち、ひらがな，漢字の部分は７１９か所あった。それ以外のカタカナおよび英字，記号の部分は、文字種の性質から誤入力部分を含む文字列が指摘される場合がほとんどなので除外した。これらを原因別に分類した結果を表１に示す。

原因が不明なものが１件あった。これは日本語としてまったく意味をなさない表現であり、誤入力であると推定されるが、何が正しいかが推測できないものである。論文誌に掲載された原稿も実験用データと同じであり、どういう誤入力かが判断できないため、他の誤入力とは別に分類した。

ここでは、技術文書を書く場合の文としてそのままで正しくはないが、人間が読んだ場合容易に意味が理解できる表現を、特殊な表現として分類する。特に書き方では、「実（■）例」などがそれにあたる。

形態素解析は１文ごとに行うため、形態素解析の前処理として表層から文を抽出する処理が必要となる。その際に表層情報を利用して処理を行なう。例えば、

普通スペースは意図的に入れるもので、一律に取ってしまうことはできない。しかし、インデントや左余白の機能を使用せず、スペースを利用して字下げを行なうなど、レイアウトに使用しているときには、文の途中に不必要なスペースがはいることになり、無視しなければならない。このような処理の不足のために起きたものを前処理不足に分類する。

表1　形態素解析による誤り指摘の原因別分類

原因		個数	百分率（%）
未登録語	外来語・英語	56	8
	固有名詞	127	18
	専門用語	54	7
特殊な表現	書き方	13	2
	口語・文語	8	1
用例など		172	24
誤入力		137	19
処理不足	前処理不足	23	17
	メタ処理不足	124	3
	能力不足	4	1
不明		1	0
合計		719	100

さらに、もう少しメタなレベルでの処理、例えば、文節の途中に括弧でくくった挿入句がある場合（例：「最適（再利用性が高いこと）[redacted]」）や箇条書き部分の前後にある不完全な文（例：「[redacted]二種類が存在する。」）を文として完成させる処理を行なっていないために起きたものをメタ処理不足に分類する。

この結果からわかるように、形態素解析の能力が不足なために誤って指摘したのは1％にすぎず、現状の形態素解析は誤り指摘にもうまく対応しているということができる。

しかし、文としては正しいが、処理の能力不足のために誤って指摘した部分も若干あった。これは1字語の合成は認めず、解析結果から棄却してしまうなどの原因によるものである。これについては、そのような規則を検討し、より詳細な情報から判断できるようにする必要がある。また、前処理・メタ処理の充実や辞書の拡充がこれからの課題である。

6．誤入力部分の評価

さらに、誤入力が原因で指摘された部分について、脱字，余字（余分な文字がはいっていること），誤字に分類する。さらに誤字については、誤入力の位置により、タイプミスと漢字入力誤りとに分ける。

全部で137か所の誤入力部分のうち、それぞれに分類した数は、表2に示すようになる。これらのうち漢字部分に当たるものは全体の4分の1程度であったが、これは、漢字以外の部分にくらべて、入力時にオペレータが誤入力に気がつきやすいためと思われる。それぞれの誤入力指摘の典型的な例を表3に示す。

それぞれの分類において、指摘された部分が実際の誤入力部分をどれだけ的確に指摘しているかを、表４に示す。これは、指摘された部分と実際の誤入力部分との近さで示す。脱字部分は脱落した文字をはさむ文字を誤入力部分と考える。

表２　誤入力の分類

分類		個数	百分率（%）
脱字		２８	２０
余字		２１	１５
誤字	タイプミス	５４	４０
	漢字入力誤り	３４	２５
合計		１３７	１００

これによれば、誤入力部分の指摘については、８５％が的確な部分を指摘している。隣接している部分を含めると９７％となる。また、指摘部分がずれている場合でも、最高で２文字であり、これらの指摘から誤入力部分を見つけることは困難ではないと考えられる。

表２　誤入力指摘の例

脱字		小型のコンピュータで翻訳を実現[]には、 る ワークステーション用[]新しいプログラミング環境を を
余字		議論の余地[]を残している。 信号に関しては合意が得らわれるが、
誤字	タイプミス	高機能ワークステーションN[]とでは多様な用途に利用可能な WS 打鍵のシーケンスが[]とくる可能性がある。 て
	漢字入力誤り	筆者等の[]する方式について報告する。 提案 自由に使用できる[]構造となっていて、 柔軟

表４　誤入力指摘の的確さ

（）内は百分率(%)

分類		含む／含まれる	隣接している	離れている
脱字		２４（８６）	３（１１）	１（　３）
余字		１８（８６）	２（　９）	１（　５）
誤字	タイプミス	４５（８３）	８（１５）	１（　２）
	漢字入力誤り	２９（８５）	４（１２）	１（　３）
合計		１１６（８５）	１７（１２）	４（　３）

７．おわりに

今回の実験結果から、形態素解析による誤り指摘に関しては充分に満足のできる結果が得られた。正しいのに誤って指摘した部分は、文自体が非文か辞書登録だけで対応できるものが大半であり、誤入力の指摘も、誤入力部分の充分近くを指摘している。今後、処理不足が原因で誤って指摘した部分については、それらの処理の充実および拡張、単語が不足しているために指摘された部分については、辞書の拡充が必要である。

現状では文章の表面的な誤りを指摘するだけであるが、文章の校正だけでなく、作成も支援するシステム、つまり、文章のあいまいさなどを除去した、わかりやすい文章の作成を支援するシステムも、今後必要となってくると思われる。

意味解析を完璧に行なうことは現在の技術では、実際的でない。特に、文章の内容を限定しない一般文に対しては不可能に近い。ターゲット別の機能を絞った局所意味解析とのかねあいで強化し、より使いやすく実用的なシステムにしていくことがこれからの課題である。

［参考文献］

1) 小山他：「文章作成支援システムの機能」
情報処理学会第３５回（昭和６２年後期）全国大会４Ｓ－４，ｐ１２５９
2) 小林他：「文章作成支援システムにおける日本語処理－文章誤りの検出－」
情報処理学会第３５回（昭和６２年後期）全国大会４Ｓ－５，ｐ１２６１