自 然 言 語 処 理 99－8
（1994．1．21）

外国人の書いた日本語文の訂正に関する基礎的研究

坂倉守 山村毅 佐川雄二 大西昇 杉江昇

名古屋大学工学部情報工学科

外国人の書いた日本語の文章の中には、日本語として不自然であったり誤っていたりするものが存在する。これらの文章を訂正する場合、従来の日本人が書いた文章に対する訂正方法とは別の訂正方法が必要であると思われる。本研究では、このような文章に含まれる誤りには具体的にどのようなものがあるかを調べ、分類を行なった。そして、分類した誤りのタイプのそれぞれに対して、文のレベルでの訂正方法を考案した。

Basic Research for Correcting Japanese Sentences Written by a Foreigner

Mamoru Sakakura, Tsuyoshi Yamamura, Yuji Sagawa, Noboru Ohnishi and Noboru Sugie

Department of Information Engineering, Faculty of Engineering, Nagoya University

When we read Japanese sentences written by a foreigner,we often feel then somehow strange. To correct such strange sentences,we think we need a method that is different from a correcting method for Japanese sentences written by native Japanese. In this paper,we examine and classify the errors in these strange sentences. And for each class of errors,we design a method for correcting errors in one sentence.

1 はじめに

近年の日本の科学技術の発展にともない、海外から学びに来る人(留学生)の数が多くなってきている。そういう人たちにとって一番の障害となるのは、日本語の習得であると思われる。一通りの日本語教育を受けてはいるものの、彼らの書く日本語の文章は、日本語として不自然であったり、誤っていたりすることが多い。日本語が上手な人が身近にいる場合は、その人に相談することで誤りを正すことができるが、必ずしもすべての留学生がそのような恵まれた環境にいるわけではない。そういった人にとって、自らの日本語を訂正してくれる文書校正システムは極めて有用であると思われる。

文書校正に関しては、これまで主に新聞・図書などの出版の分野で開発・実現が行なわれてきた。これらにおいては、漢字の変換誤りや誤字・脱字などの「入力誤り」をその校正対象としていた。日本人の書く文章を校正する場合には、このようなタイプの誤りだけを考えれば十分かも知れない。しかし、外国人の書く文章の場合には、「入力誤り」だけでなく、文書生成時の「生成誤り」を考慮する必要がある。

そこで本研究では、外国人が作成した日本語文を対象に、その誤りのタイプの分類を行なう。そして、分類された誤りのそれぞれに対して、その訂正の方法を考案する。

2 生成誤りの分類

通常、文は、ある発意の内容から「語彙選択」や「文法規則適用」などを経て生成される。この際、不適切な単語を選択したり、誤った文法規則を適用したりすれば、不自然な文あるいは誤った文が生成されることになる。このような誤りを我々は、「生成誤り」と呼ぶことにする。この章では、この「生成誤り」に具体的にどのようなものがあるかを調べるために、我々の研究室に所属する中国人留学生が日常的に作成した日本語文のうち、誤りを含んでいると思われる 134 個の文を調査し、そこに含まれる 173 個の誤りのタイプの分類を行なった。

この結果を表 1 に示す。

表 1: 誤りの分類結果

誤りのタイプ	誤り数
サ変動詞の名詞化時の誤り	10
数詞の使用における誤り	7
母国語の干渉による語順の誤り	7
取り立て詞の使用誤り	7
時制の誤り	5
自動詞・他動詞の誤り	8
接続の誤り	15
題述の誤り	10
漢語選択誤り	8
その他	96

表に分類してあるのは、誤り数が 5 個以上あったものである。誤り数が 4 個以下のものは「その他」に分類した。

以下、それぞれの誤りのタイプについて詳しく説明する。

2.1 サ変動詞の名詞化時の誤り

名詞の修飾節内において、連体修飾形にしなければならない修飾句が連用修飾形になっているものである。この誤りが発生する場合、被修飾の名詞は例外なくサ変名詞である。例 1 では、下線部の句の後に助詞「の」が必要である。

(例 1) Photometric Stereo & Active Observer から三次元構造の推定は…

一般に、サ変名詞句は、対応するサ変動詞句を名詞化することによって生成できるが、この際、元のサ変動詞から「する」を取り除き、その連用句を連体形にする必要がある。すなわち、

(名詞化規則) 連用句+サ変名詞+「する」
→ 連体句 + サ変名詞

この規則を誤って記憶していたために、この誤りが生じてしまったと考えられる。

2.2 数詞の使用における誤り

日本語では、数詞による名詞の修飾は、数詞+「の」+名詞の形になるが、これを、数詞+名詞の形にしてしまったものである。

(例 2) まず、二つ modulesの関係…

通常、数詞は、数+助数詞から構成され、これは名詞の扱いとなる。すなわち、

(数詞規則) 数+助数詞 → 数詞 (名詞)

これを、名詞ではなく形容詞の扱いにしてしまったために生じた誤りであると考えられる。

2.3 母国語の干渉による語順の誤り

日本語と英語・中国語等の他の言語との主な語順の違いは、主動詞の文中での位置等であるが、この他にも細かいところでいくつかの違いがある。

(例 3) 物体の textured 表面或は textureless 表面に対するいずれも適用できる。

中国語では、「A あるいは B のいずれに対しても」は、「A あるいは B に対するいずれも」という語順になる。このような細かな語順の違いを知らなかったために生じてしまった誤りであると考えられる。

2.4 取り立て詞の使用誤り

「は」による取り立ては、外国人が日本語を学習する上で特に問題になるものである。「取り立て」を行なってはいけないところで「は」を用いているものと、「取り立て」を行なうべきところで「は」を用いていないものとに分けることができる。

(例 4) 条件に関わらず物体の Shape を抽出することができる方法、これは最もよい方法と思います。

(例 5) 人間は最も優れると言っても、何時でも物体を認識することができないが…

「は」と「が」がどのような規則で用いられているのかを正確に記述することは、我々日本人にとっても容易なことではない。また、必ずしも文のレベルで訂正できるものではないので、この誤りを訂正するには、文を越えた文脈のレベルの処理が必要になる。本稿ではこの種の誤り訂正は取り扱わないことにする。

2.5 時制の誤り

一つの文の中の並列節で時制が一致していないもの等、不適切な時制が使用されているものである。文脈のレベルの処理が必要なものもあるが、本稿では取り扱わない。

(例 6) float data の場合には、予想と一致しましたが、int data の場合には、予想とちょっと違います。

この種の文は、話し言葉では容認されることもある。

2.6 自動詞・他動詞の誤り

自動詞と他動詞を混同してしまったタイプの誤りである。例えば、次の例 7 では、「たくさんの情報を集めて」と他動詞を用いて記述すべきところを、「たくさんの情報を集まって」と自動詞を用いて記述している。これは、「集める」と「集まる」とが表層的に良く似ているため、混同してしまったものと考えられる (日本語では、自動詞とそれに対応する他動詞とは、表層的に一文字だけ異なっているものが多い)。

(例 7) たくさんの情報を集まって、分析して認識する。

次の例 8 は、下線部分だけを見るのならば、誤りはない。しかし、先行する文「その p,q を解けば」があるので、下線部は、「面の傾きが求められる」のように自発表現を用いるか、「面の傾きを求めることができる」のように可能表現を用いなければならない。

(例 8) その p,q を解けば、面の傾きを求める。

この誤りは、正確には例 7 のような単純な自動詞・他動詞の混同ではなく、自発表現の選択誤り (自発表現すべきところで自発表現していない、あるいはその逆) であるので、同じタイプの誤りに分類するのは不自然かもしれない。しかし、例 7 で「たくさんの情報を集めて」と訂正するのは自発表現の選択誤りの発生を防ぐためである (「たくさんの情報が集まって」のように訂正すると自発表現の選択誤りになる) とも考えられるため、これらを同じタイプの誤りとして分類した。

2.7 接続の誤り

不適切な接続語句 (接続詞や接続助詞) の使用誤りである。これには、余分な (重複した) 接続語句の使用や接続語句の不適切な省略も含まれる。

(例 9) もし、カメラが物体の真上にあるときは、カメラの影が物体に当たってしまう可能性がある。

この例は、「もし」と「ときは」を重複して用いているので (あるいは「もし」を用いながら「ならば」のような仮定形で終っていないので)、不自然になっている。「カメラが物体の真上にあるときは」あるいは「もし、カメラを物体の真上に配置してしまえば」のように訂正しなければならない。

2.8 題述の誤り

題目と題述が呼応していない (題述が題目を叙述していない) タイプの誤りである。次の例 10 は、「…方法は」と題目を提示しておきながら、「…する」のように題述しているので不自然である。「…することである」のように記述すべきである。

(例 10) この問題を避ける一つの方法は、先に、segment してから、shape from shading する。

この種の誤りは、日本人でも、特に文章が長くなった時に、起こしやすい誤りである。

2.9 漢語選択誤り

日本語で使われている言葉の中には、古く中国から輸入されたものが数多く存在する。それらの中には、長い年月を経たために、形態 (漢字) が異なるものや (漢字が同じでも) 意味が異なるものなどがある。次の例 11 の場合、日本語では通常、「影」を用いる。

(例 11) 回りの誤差は、光源の蔭のためと思います。

この種の誤りは、漢字を常用する中国人に特有のものであって、漢字を用いない欧米人には見られないものであると考えられる。

2.10 その他の誤り

以上述べてきたものの他に、敬語の使用誤りや、送りがなや濁点・半濁点の誤りのような表記上の誤りがあった。濁点・半濁点の誤りは、例えば、「進んでいない」→「進んていない」のように濁点が消失あるいは付加してしまうもののことで、表記上の誤りとして比較的多く現れる。日本語と母国語との音 (母音や子音) の違いに起因するものであると考えられる。

3 文の訂正

2 章では誤りの分類を行なった。ここではそれらの内のいくつかについて、その訂正の方法を考案する。

3.1 サ変動詞の名詞化時の誤りの訂正

この種の誤りは基本的に構文解析ルールにより検出できる。

寺村 [2] によると、動詞の名詞化時の、連用補語 (名詞+格助詞) から連体形 (名詞+接続助詞) への転換にさいして働くルールは、

(イ) 格助詞が「ガ、ヲ」の場合は、それを消去して「ノ」を付ける

(ロ) 格助詞が「ヘ、デ、カラ、マデ、ト」の場合は、それを残してその後に「ノ」を付ける。

(ハ) 格助詞が「ニ」のときは、その意味によって、その意味に近い「ヘ、デ、カラ」で代行させる。ただし、時を表す「ニ」の場合に限り、それを「の」と入れ替える。

(ニ) 比較の基準を表す「ヨリ」は連体化できない。

とある。

そこで、この種の誤りが検出された場合、名詞+格助詞の形であれば、上の (イ) ～ (ニ) の内、格助詞に応じた規則を使用して書き換えればよい。

3.1.1 文脈自由文法を用いた誤りの検出・訂正

文脈自由文法を用いて構文解析を行なう場合の誤りの検出・訂正の方法を考えてみる。

サ変名詞に関する文法規則が以下のようであるとする。

名詞句 → 連体補語+サ変名詞

連体補語 → 名詞+格助詞「へ、で、から、まで、と」+「の」

連体補語 → 名詞+「の」

この文法で構文解析すると、先の例1では、下線部が連体補語でないと判断されるのでうまく構文解析できない(誤りが検出される)。そこで、次の文法規則を加えて再び構文解析を行なう。

連体補語 → 名詞+格助詞

うまく構文解析できた場合、上の文法規則が適用された場所が存在するので、その場所を調べる。以下のような書き換え規則を用意し、調べた場所にある格助詞に、適切な規則を使用して格助詞を書き換える。

「が」 → 「の」
「を」 → 「の」
「へ」 → 「への」
「で」 → 「での」
「から」 → 「からの」
「まで」 → 「までの」
「と」 → 「との」
「に」 → 「への」
「に」 → 「での」
「に」 → 「からの」
「に」 → 「の」

例1の場合は「から」→「からの」を適用することになり、うまく訂正できる。

3.2 数詞の使用における誤りの訂正

この種の誤りは、基本的に構文解析ルールによって検出できる。訂正するには、

数+助数詞+名詞 → 数+助数詞+「の」+名詞

の書き換え規則を適用すれば良い。

3.3 母国語の干渉による語順の誤りの訂正

母国語の干渉による語順の誤りは、詳細な構文規則を作成して検出・訂正しても良いが、それよりも、ある特定の表現との整合をとって書換えを行なう方が容易である。例えば、先の例3では、

名詞句1+「あるいは」+名詞句2
+「対するいずれも」
→ 名詞句1+「あるいは」+名詞句2
+「いずれに対しても」

のような書き換え規則を用意して、書き換えればよい。この処理は極めて単純に行なえるが、反面、柔軟性に欠ける。このため、実用化に際しては大規模な調査を行なって、この種の書き換え規則をできるだけ多く作成しておく必要がある。

3.4 時制の誤りの訂正

並列句の時制の不一致などの不適切な時制の使用は、構文解析結果で時制が適切かどうかを調べ、不適切な場合には適切な時制に訂正する。

3.5 自動詞・他動詞の誤りの訂正

単純な自動詞・他動詞の混同は、動詞の表層格フレームを用いることにより抽出できる。その訂正は、自動詞と他動詞を置き換えたり、格助詞「を」と「が」を置き換えたりすることにより実現できる。この際どちらを行なうかは、概ね、前後の文の主動詞の自他によって決めれば良い。

自発表現の選択誤りも、同様に前後の文の主動詞の自他との一致で概ね検出・訂正ができるが、例8のように、接続関係によって制御されることもあるので、接続関係を考慮した自発表現の選択規則を作成する必要がある。

3.6 接続の誤りの訂正

この種の誤りは、基本的に構文解析ルールによって検出できる。訂正に関しては、重複接続語句、不適切省略のいずれに対しても、動詞間の意味関係を利用する方法が考えられる。

3.7 題述の誤りの訂正

題述で、主に「…は…することである」のようにすべきところを「…は…する」のように誤ってしまうものであるから、「…は」にあたる部分が主動詞の格 (主として動作主格) として適切であるかどうかをその深層格フレームを用いてチェックし、不適切な場合には、その末尾に「ことである」を付加すればよい。例えば、例 10 では、「方法」は、「する」の主体として不適切であるから、その末尾に「ことである」を付加することにより訂正できる。

3.8 漢語選択誤りの訂正

この誤りは、語彙上の問題であるため、検出・訂正には、日本語の語彙辞書と中国語の語彙辞書が必要である。日本語の語彙辞書で辞書引きした時に未登録語となってしまうものは中国語の語彙辞書で辞書引きし、対応する適切な日本語の単語に置き換えれば良い。

4 考察

以上の結果から、構文レベルで処理できる誤りについては、この方法で訂正することができると思われる。

さらに、意味レベルでの処理が必要なもののうち、いくつかについては、意味情報などの利用によりこの方法で訂正できると思われる。

しかし、問題点もいくつか残っている。

ひとつは、分類に使用した文の数が少ないことである。絶対数が少ないため、誤りのタイプとして分類できそうでありながら「その他」に分類した誤りがいくつかあった。文の数を増やせば、そのような誤りが、誤りのタイプの一つとして分類できるかもしれない。逆に、現在誤りのタイプとして分類されているものも、たまたま数が多かっただけで実際には妥当な分類でない可能性もある。

ふたつめは、誤りのタイプごとの処理であるため、全ての誤りを今回述べたような方法で訂正しようとすると文法の数が非常に大きくなることが予想されることである。そのため、このままでは実用化は困難であるが、河合ら [1] の構文標準化法などを使用することによって対処することができると思われる。

さらに、「その他」に分類されている誤りが多いことも問題である。今の分類では、「その他」の数が全誤りの半分を越えているため、たとえ、「その他」以外の誤りの全てを処理できたとしても、誤りを含んだ文の内、半数も処理できないと思われる。

今後の課題としては、

1 分類に使用する文の数を増やし、分類の信頼性をあげるとともに、「その他」の分類をできるだけ少なくする。
2 今回細かく触れなかった誤りのタイプについても訂正に関する規則を明確にし、実際のシステムを作成できるようにする。
3 今回分類に使用した文は中国人留学生のものだけであるため、他の国からの留学生の文についても分類を行なう。
4 今回取り扱わなかった文脈のレベルでの処理を行なう。

などがあげられる。

5 まとめ

本研究では、外国人が作成した日本語文を対象に、その誤りのタイプの分類を行なった。そして、分類された誤りのそれぞれに対して、その訂正の方法を考案した。今後は、分類に使用する文の数を増やし、分類の信頼性をあげていくとともに、「その他」の分類をできるだけ少なくしていく予定である。

参考文献

[1] 河合敦夫、杉原厚吉、杉江昇：“英文の誤り検出のための構文標準化法”
電子通信学会論文誌 Vol.J-66D, No.5, pp.511-518 (1983)

[2] 寺村秀夫：“日本語のシンタクスと意味 第III巻” くろしお出版 (1991)