自然言語処理 117－2
（1997. 1. 20）

外国人の書いた文章の助詞使用誤りの抽出

橋本利典　　島田静雄
埼玉大学工学部情報システム工学科

概要

外国人の書いた文章には、日本人の書いた文章とは違った文法誤りなどが存在する。彼らの文の誤りを分析することは、新たな側面から日本語を解析することにつながる。そこで本稿では、外国人の書いた日本語の文章を調査分類し、その中でも文法誤りである助詞の抽出・訂正方法を考案した。本手法では、助詞誤りの判断基準として動詞の選択制限規則を用いた。漢字の持つ意味情報から名詞の意味素性を推定し、その名詞に付属する助詞と動詞との親和性に着目した。

Grammatical Errors Of Using Particles in Japanese Sentences Written by Foreigners

Toshinori HASHIMOTO,Shizuo SHIMADA
Department of Information and Computer Sciences,Saitama University

In Japanese sentences written by foreigners, there are grammatical errors and strange expression different from those written by native Japanese. Analyzing such errors,we can anlyse Japanese sentences even if they are written by native speakers. In this paper, we investigate and classify the errors in Japanese sentences written by foreigner.we develop the extraction method of the particle which contain grammatical errors in thier sentences.We describe the method to use the information of the selectional restriction of verbs. A noun with KANJI has the information of the meaning.The method estimate the relation between the particle connecting nouns and verbs.

1　はじめに

最近の日本語ワードプロセッサには、文書校正支援システムが付属しているものが多くなってきた。ユーザーからの計算機による文章の誤りの自動検出や校正支援に関する要求が高まっているからである。しかし、現在の校正支援システムは、校正の対象が日本の国文法に基づいて書かれた文章、つまり日本人の書いた文章である。それゆえ、その文法規範から外れている文章には対処しきれていない。また、話し言葉など文法の範囲を越えたものを人間は理解できるが、システムには理解できないものが多い。本当に実用的なシステムとなるためには、現在では対象外の文章であっても将来的には処理可能にする必要がある。

さて、外国人が書いた日本語の文章には、誤字・脱字などの表記上の誤りや文法上の誤りのある文が混じると同時に、文法上は正しいが日本人から見て意味の通じない文がある。彼らの書いた文章を分析するには、日本の国文法とは違った角度から日本語を解析しなければならない。彼らの書いた文章の中から誤りを抽出するシステムの構築は、新たに日本語の構造を理解することになる。また新たな側面から日本語を解析することは、「コンピュータで扱える、より柔軟な文法規則」を作り出す。将来的に日本語処理技術の向上になり、より実用的な機械翻訳・文書校正システムに役立つと考えている。

そこで、本研究では、まず外国人の書いた文章から誤り調査分類を行なう。次に日本人が書いた文章には見られない誤りを分析する。外国人特有の誤りである助詞誤りの抽出に関する文法規則を考案し、その規則の有効性の評価として助詞の使用誤りの抽出を行なう。

2　誤りの調査と分類

2.1　サンプル調査

外国人の書いた日本語文章の中から、誤りの調査を行なった。サンプル文書のデータは、表1に示す。サンプル文書の形式は、論文、手紙文、感想文のいづれかである。手書きで書かれたものをキーボードから入力し、テキストデータにしたものである。サンプル提供者の総人数の内、漢字使用語圏の人が３５名と非漢字使用語圏の人は、２５名である。一文の区切りは、句点から句点までとし、平均文長は、総文字数を文数で割ったもの、漢字使用率は、総文字数中の漢字の割合を示し、ひらがな、カタカナ、英数・記号の使用率についても同様に算出した。また、学習程度の度合として使用漢字の種別を表2に示す。使用漢字の種類を学習漢字１～６年生、常用漢字、および上記以外の第一水準漢字、第二水準漢字の割合を調べた。

表1　サンプルデータ

文字情報	数
サンプル文書数	99 文書
総人数	60 名
総文字数	49363 文字
総文数	1579 文
平均文長	31.2 文字
漢字使用率	28.6 %
ひらがな使用率	50.2 %
カタカナ使用率	10.2 %
英数・記号使用率	12.0 %

表2　使用漢字の種別

漢字の種別	割合
学習漢字１年生	25.73 %
学習漢字２年生	24.44 %
学習漢字３年生	13.73 %
学習漢字４年生	9.67 %
学習漢字５年生	12.44 %
学習漢字６年生	8.24 %
常用漢字	5.31 %
上記以外の第一水準漢字	0.40 %
第二水準漢字	0.05 %

表１、表２の結果より、平均文長が約３０字であることからやや文が短く、小学生程度、漢字使用率が約３０％であり、学習・常用漢字の使用率などから中学生程度と見ることができ[1]、全体的に小学生高学年から中学生程度の文章であることがわかる。

2.2　誤り分類

次にサンプルから日本人が見て誤りと判断したものの分類を表3　に示す。1579 文中に 712 文に誤りが見つかった。総誤り数は、883 個で

表 3 誤り分類結果

誤りの種類	誤りの数（個）	割合（%）
1:助詞の誤り	303	34.3
2:動詞の誤り	185	21.0
3:名詞の誤り	174	19.7
4:形容詞の誤り	81	9.1
5:副詞・接続詞の誤り	52	5.9
6:その他	88	10.0
7:合計	883	100.0

あるが、同一の文に何度も誤りが見られる場合もあるので、誤り文数より多くなっている。誤りの種類は、品詞別に大まかに分けたものである。以上の結果、助詞の誤りが全体の 34.3 %を占めて、もっとも多いことがわかる。助詞の誤りは、日本人が書いた文章には見ることができないものがほとんどであった。次に助詞誤りの詳細を表 4 に示す。

2.2.1 助詞誤りの分類

表 4 助詞誤り分類結果

誤りの種類	誤りの数（個）	割合（%）
1:使用・選択誤り	163	53.7
2:欠如	86	28.4
3:不必要	30	10.0
4:表記上の誤り	17	5.6
5:その他	7	2.3

助詞の使用・選択誤りが反数を占めている。これらの誤りの判断基準となったものは、述語である。また助詞の欠如においても、名詞と名詞を接続する「の」の欠如以外、不必要な助詞においても、同様に述語で判断できた。しかし、一文からは判断できず、文脈上から判断したものも小数だが含まれている。表記上の誤りは、誤字や濁音など音に関するものである。上記以外は、その他に含まれている。

3 助詞誤りの抽出手法

3.1 助詞使用誤りの判断

我々は、助詞の使用が誤っていることを判断するには、述語を基準している。我々は、述語から文を理解しているのである。次の例を用いて説明をする。

- 例「私は日本に興味が持っています。」

「私は日本に興味が」までを読んだ時点では、助詞の誤りを判断できないが「持っています。」を読むと、「興味が」の「が」が誤っていると判断できる。また、「私は日本に興味が」の後に続く動詞を予測することができる。例えば「ある。」などである。これは、述語である動詞から共起しうる名詞句、修飾句をかなり正確に予測できるからである。しかし、目的語が必要な他動詞の「持つ」を目的語の必要のない自動詞「ある」に訂正すればよいということも言えるが、他動詞と自動詞を変換すると作者の意図する内容を大きく変化させてしまう。よって、この本稿では、動詞の選択誤りでなく助詞の誤りに注目し、その誤りの判断基準を動詞とする。

3.2 動詞辞書

動詞には、それぞれ選択制限規則が存在する。動詞には親和性の良い名詞があり、それに接続する助詞にも制限がある。意味解析で用いられる格フレーム (case frame)[2]の概念である。そこで、動詞と助詞との制限規則をどのように行なうかを列挙する。動詞の選択制限規則には、親和性のある名詞に対して意 味的な制限を持っている。名詞に対しての制限を仮に「意味素性」とする。[34]

- 動詞を自動詞・他動詞に分類
- 動詞それぞれに親和性のある助詞と名詞の基本文型の作成
- 名詞の意味素性のみを付加することで助詞の意味情報も兼用

計算機用日本語基本動詞辞書IPAL(Basic Verbs)[3]から上記の条件を満たす情報を取り出し、新たに辞書を編集する。

・動詞辞書

動詞辞書は、基本的な単漢字動詞８００語を用いる。使用頻度の低いものは省いた。見出しに動詞漢字、動詞の読み、語幹の読み、基本文型、

接続可能助詞、それに親和性のある名詞の意味素性を付加している。
意味素性の分類については、次節で簡単に説明をする。動詞辞書の例を以下に示す。

[例] :「持つ／もつ／も／N1 が N2 を／がをにと／N1：HUM、ORG、／N2：PRO、MEL、LOC」

3.3 漢字辞書

動詞辞書に名詞の意味素性を加えることによって、親和性のある助詞の限定を行なうことができるが、すべての名詞に意味素性を付加するのには膨大な作業が必要である。そこで、単漢字に着目し、意味素性を与え、２字以上の熟語についてはその組合せによって意味素性を推定する方法を取る。
使用される漢字の頻度から、学習漢字１～６年生の１００６字に１６種の意味素性を与える。意味素性の分類は、IPAL 名詞辞書[4]で使用されているものを採用する。その例を表 5 に示す。

表 5 意味素性分類

略号	素性名	例
ANI	動物	犬・虫・貝
HUM	人間	女・男・人
ORG	組織・機関	国・郷・県
PLA	植物	花・森・木
PAR	生物の部分	頭・足・腕
NAT	自然物	山・川・石
PRO	生産物	紙・車・布
PHE	現象	光・音・風
ACT	動作・作用	行・立・入
MEN	精神	心・悩・気
CHR	性質	美・青・赤
STA	状態	安・静・固
REL	関係	縁・連・系
LOC	空間・方角	場・外・右
TIM	時間	時・日・夕
QUA	数量	数・人・巻

表 5 以外に種別としては、動詞漢字・形容詞漢字・名詞漢字・副詞漢字・接辞漢字と分けた[5]。それぞれの特徴として、

- 学習漢字の中に主に動詞として使われる漢字は 399 字ある。これらを動詞漢字と呼ぶ。意味素性は、ACT（動作･作用）または PHE（現象）である。動詞漢字が２字合わさった熟語は、ACT または PHE となる。

- 学習漢字の中に主に形容詞として使われる漢字は 127 字ある。これを形容詞漢字と呼ぶ。意味素性は、CHA（性質）または STA（状態）である。

- 学習漢字の中に主に名詞として使われる漢字は 459 字ある。これを名詞漢字と呼ぶ。意味素性は、ACT･CHA･STA 以外である。

- 学習漢字の中に主に副詞として使われる漢字は 3 字ある。これを副詞漢字と呼ぶ。意味素性は、STA である。

- 学習漢字の中に主に接辞詞として使われる漢字は 18 字ある。これらを接辞漢字と呼ぶ。意味素性は、CHA または STA である。

このように分類したが、単漢字で音読みする場合と訓読みする場合とでは意味素性が異なる場合がある。その場合複数の意味素性を持つとする。（最大４）

・漢字辞書

漢字辞書は、学習漢字１００６語を用いる。見出しに単漢字、意味素性１、意味素性２意味素性３、意味素性４、種別、音読み、訓読みを用いる。意味素性の優先順位は、１から４の順である。

3.4 意味素性の推定

文章中で単独に一字の漢字で使われている場合、優先順位の高い意味素性１を用いて推定する。単漢字で音読みする場合と訓読みする場合とでは、意味素性が異なる場合がある。例えば接尾語として数詞となる漢字などは、その直前の文字が数字であれば QUA（数量）の素性を持つとする。

1. 「月」:「つき」,「がつ」,「げつ」,意味素性1[NAT],意味素性2[TIM],意味素性3[QUA]
2. 「月」→「つき」 :NAT,TIM
3. 「１月」→「いちがつ」 :QUA,TIM

また２字の熟語は、次のように推定する。漢字同士に意味的なつながりを持っている。そのような観点から、後半にくる漢字（後半漢字）がその熟語の主な意味を成していることがわかる。その特性を考慮し熟語の意味素性を決定する規則を下記に示す。

- 基本的に２字の熟語の場合、前半漢字が後半漢字を修飾し説明を加えているものとする。この場合、主な意味素性は後半漢字のものとする。
- 前半漢字および後半漢字に意味素性が複数あり、それぞれ同じ意味素性を持っているならそれを主な意味素性とする。

２字の熟語の意味素性の推定例を次に示す。

1. 「勉」＋「強」→「勉強」
2. 「勉／ ACT ／動詞漢字」 および 「強／ CHA ／ ACT ／形容詞漢字」
3. ACT : ACT = ACT
4. 「勉強」= ACT

「勉」は、「勉める」という動詞漢字として位置付け、ACT が意味素性となる。「強」は、「強い」という形容詞漢字として位置付け、CHA が主な意味素性となるが、「強いる」という動詞の使われ方をする場合もあるので ACT も意味素性としてなっている。この二つの漢字が合わさった場合、上記の規則により 3 のように推定し、「勉強」の意味素性は、ACT とする。「勉強」は「勉強する」という「する名詞」になっていることからも、行動を表していることがわかる。

３字以上は、２字の場合と同様とし、一番最後にくる漢字の意味素性を主な意味素性とする。固有名詞は漢字同士のつながりがないものが多く、以上の規則に当てはまらない。また、ひらがな語・カタカナ語も単語ごとに意味素性を指定しなければならない。しかし、本稿では取り扱わないことにする。

3.5 助詞の誤り抽出

動詞辞書・漢字辞書を活用し、以下の手順で行なう。

1:動詞を取り出す。
2:動詞の直前の語を取り出す。
3:助詞であれば、次の処理 4 を行なう。
ルール 1 に当てはまれば、その直前に助詞がある可能性が高く、さらにその直前の語を取り出し、処理 3 を再度行なう。もし名詞であれば、助詞が欠落しているものとし、誤り候補として抽出する。
4:動詞辞書から情報を読み出す。
接続可能助詞と名詞の意味素性を読み出す。
5:助詞制限規則と助詞候補を比較する。
ルール 2 を適用し、接続可能な助詞でないなら、誤り候補として抽出する。
6:助詞の直前に名詞がある場合
単漢字に区切り、名詞辞書を読み出す。名詞が熟語である場合、熟語の意味素性を決定する。
7:意味素性を比較する。
もし、意味素性が異なる場合、誤り候補として抽出する。

○ルール 1
助詞以外で動詞の直前に接続するのは、読点・副詞・形容詞・形容動詞・数詞である。
○ルール 2
・格助詞「を」は、必ず他動詞の目的語に接続
・態（ヴォイス）により、動詞の基本文型と異なる助詞の接続可能
・他動詞に接続する場所を示す「に」は、「へ」と置換可能、その逆も可

4 評価

4.1 評価手順

前章で提案した助詞誤りの抽出手法の精度を検証するために、前章の手順で実験を行なった。対象文書としてサンプル文書から抜粋した 80 文（Test1）と、作為的に誤りを持たせた 50 文（Test2）、および新聞からの全く誤りのない 50 文（Test3）を用いた。また、本稿で使用する誤り抽出精度として次の式を定義する。再現率は、抽出もれのなさを示し、不適合率は、無駄な抽出の度合を示す。

$$\text{再現率} = \frac{\text{正しく抽出できた助詞の数}}{\text{抽出すべき助詞の数}} * 100$$

$$\text{不適合率} = \frac{\text{誤って抽出した助詞の数}}{\text{抽出した全ての助詞の数}} * 100$$

4.2 結果と考察

評価結果を次の表 6 に示す。

表 6 評価結果

	再現率 (%)	不適合率 (%)
Test1	66.3	3.4
Test2	88.3	0.2
Test3	― ― ―	100.0

表 6 では、Test1、Test2 の不適合率が低い値を示しているので無駄な抽出が少ないことを表している。Test3 の再現率に数値がないのは、抽出すべき助詞の誤りがないためである。新聞の文章には誤りが含まれていないことを表している。しかし、不適合率が 100 %を示しているため、誤って抽出してしまったものがあることを表している。また、Test1 では、再現率が少し低い値を示しているので、抽出もれが起こっていることを表している。その原因を分類した結果を以下に示す。

Type A 助詞制限に当てはまる範囲での誤り。これは、助詞が付属している名詞の意味素性と辞書の意味素性とがうまく判断されていないことに関わっていてるもの。固有名詞やカタカナ語など。（約 80 %）

Type B 助詞の順番や動詞の表現形態によって助詞限定規則が変化しているもの。（約 20 %）

5 まとめ

本稿では、外国人の書いた文章における誤りの調査分類を行ない、文法誤りである助詞の使用誤りの抽出を行なった。動詞の選択制限規則を用い、名詞の意味素性情報による助詞の使用誤りを推定する方法を述べた。その際、大規模な辞書を用いずに、単漢字の情報より名詞の意味素性の推定方法も考案した。

6 今後の課題

- 今回提案した抽出手法は、基本的に動詞の直前に現れる助詞についてのみ行なっているが、順番が入れ替わってしまっているものや動詞から離れているものにも対処し、TypeB の誤りの抽出精度を上げる。
- 名詞の意味素性の推定方法にさらなる考慮が必要。それによって TypeA の誤りの抽出精度を上げる。
- 他の用言（形容詞・形容動詞）についても制限規則の辞書を作成する。

参考文献

1) 安本美典：「説得の文章技術」講談社現代新書 p(1983)
2) 電子情報通信学会 長尾真: 日本語情報処理 (1984)
3) 情報処理振興事業協会:計算機用日本語基本動詞辞書 IPAL (Basic Verbs) (1987)
4) 情報処理振興事業協会:計算機用日本語基本名詞辞書 IPAL (Basic Nouns) (1996)
5) 鈴木英二 中挾知延子 島田静雄 :「漢字シソーラスの構築と語句解析への応用」 情処学会第 52 回全国大会,4B-6,(1996)