8L-02

# 日本語文章校正ツール“Chanterelle” －助詞の誤り検出について－

奥村 薫
Microsoft Corporation

## 1. 始めに

助詞の用法は、日本語の中でももっとも難しいもののひとつとされている。日本語学習者にとって難しいのみならず、日本語を母国語としている人の場合でも、文章を練り直している際などに、しばしば助詞の不整合が発生し、ねじれた文になってしまうことがある。助詞の誤り総てを網羅することは現在の技術水準では困難であろう。しかし入力ミス・データベースを当たっている内に、比較的容易に検出しうるグループがあることに気づいた。

本稿では、その助詞の用法の誤り検出手法と精度について述べる。なお、当誤り検出は日本語校正ツール(コードネーム Chanterelle)上で実装されており、Word:mac2001 及び次期 Word に搭載される。

## 2. 助詞の誤りパターン

### (1)「と」の用法

「と＋が」や「と＋を」は、基本的に２者を比較する語法である。そこで、比較されるべきものが存在しているかどうかをチェックする。

基本ルールⅠ
「～と」無しに「～とが」「～とを」が出現するパターンを検出する。

**Ⅰ誤り検出文例**

- **実用車を中心に外車、国産車とを展示。**
- **大自然に溶け込むとができるリゾート・コース。(→溶け込むことが)**

### (2)「の」の用法

「の」には多くの意味があるが、原則的に体言を修飾する。本来ならば、構文解析をしなければいけないところだが、元来修飾されるべき体言が無い場合に構文解析結果を信じることができるとは限らない。むしろ形態素解析結果を利用して、直後の用言の用法を調べることにより、誤りを検出できることが多い。

基本ルールⅡ
「～の」で終わる文節の次に用言が来て、用言が、
- ⅡA：名詞を修飾できない
- ⅡB：名詞を修飾できるが、その後に名詞を修飾できない用言または文末が来る

パターンを検出する。
ここで、用言とは「名詞列＋サ変動詞」を含む。

**ⅡA 誤り検出文例**

- **部屋の取れてから、明日の時間をご連絡します。(→部屋が取れてから/部屋の予約が取れてから？)**
- **フィールドでの視点の変更するなど、(→視点を)**
- **オーストラリアの小学校のはなしましょう。(→はなしをしましょう/小学校についてはなしましょう？)**
- **彼女の置かれ環境を、(→置かれた)**

**ⅡB 誤り検出文例**

- **花は不快なにおいのする。(→においがする)**
- **ケアの要望に積極的の応答する。(→積極的な)**
- **キャラクターグッズの扱っている。(→グッズを)**

ただし、基本ルールのみでは、正しい日本語も検出されてしまうことが多い。そこで、次の修正を加える。

Ⅱ-1：さらに（補助）用言に繋がることのできる用言の場合には、次の用言を調べる。

- **ユーカラなどの歌って伝えられる物語へと展開した。**

Ⅱ-2：名詞列＋サ変動詞で、副詞的名詞・形式名詞・量的名詞・時間に関する名詞等を含む場合には許される。

- **どれくらいの時間発電しているか？**
- **普通春と冬の年二回発生する。**

---

Japanese Proofing Tool "Chanterelle" - Detecting Marker (“JOSHI”) Usage Errors
Kaoru Okumura, Natural Language Group, Microsoft Corporation
One Microsoft Way, Redmond WA 98052-6399, U.S.A. e-mail: kaoruo@microsoft.com

## 3. 再現率の考察

### （1）検出頻度

さまざまな入力ミスを集めたデータベース中、今回の「助詞の用法」ルールでどの程度、誤りを検出できるかを調べた。

| | 総入力ミス数 | 検出数 | % |
|---|---|---|---|
| DB1 | 1830 | 28 | 1.5% |
| DB2 | 280 | 6 | 2.1% |

パーセンテージで言えば少ないように見えるが、この数字は総ての入力ミスに対して占める割合であるので、1.5%～2%とは言え、かなり有効性が高いと言えよう。

### （2）検出もれの考察

過剰検出を抑制するために導入した制約によって、誤りではあるが検出されなくなった文例もある。

- **事件のおき状況を調査した結果～（→おきた）**
- **濁流はあらゆるものの飲み込み、押し流していった。（→ものを）**

連用形の動詞は、体言化しているのか用言として用いられているのかの判断が難しいため、当ルールからは除外されている。高度な解析をすれば、判定できる可能性もあるが、もともと間違っている文だけに正しい解析をするのは容易でないだろう。また、「疑わしきは罰せず」という日本語校正の指針により、過剰検出よりは過小検出ぎみに設計してある。

## 4. 適合率の考察

（1）過剰検出の頻度

さまざまなコーパスにおいて、本ルールによる過剰検出数を測定した。

| コーパス | 文字数 | 過剰検出数 | ページ(*) |
|---|---|---|---|
| 新聞 | 5, 361, 785 | 20 | 383 |
| 書籍 | 34, 448, 909 | 252 | 195 |
| Web カタログ | 3, 895, 256 | 29 | 192 |
| 小説 | 11, 596, 682 | 282 | 59 |
| 週刊誌 | 875, 374 | 19 | 66 |

ページ(*)は、文庫本（1 ページ約 700 文字と仮定）として換算した際に、平均何ページに一個検出されたかを表す。

くだけた表現や方言、古い日本語、非標準的な文章が多いほど検出数が多いことが観察されている。

また、総検出数と正しい検出の比率の一例を次にあげる。

| コーパス | 総検出数 | 正検出 | % |
|---|---|---|---|
| Web カタログ | 64 | 35 | 55% |
| 新聞 | 35 | 15 | 43% |

これら 2 つのコーパスは、共に校閲済みのものであり、原則的に間違いが無いはずであることを考慮すれば、約半数が正しい検出であるとはかなり良い成果であるといえる。実際校閲前や、書きかけの文章では、殆どが正しい検出となる。

### （2）過剰検出の考察

当然ながら、過剰検出の多くが形態素解析の誤りに起因するものである。

**未知語による例:**

- **あの夜のみそかごとを今にも人が知って～**

「みそかごと」が未知語であるため、「みそ（味噌）かご（籠）とを」と解釈し、「～とを」ルールでチェックされている。

**形態素解析エラーによる例:**

- **頭のわるかお人たいのう。**
- **当節のされちまうから。**

「頭の/わる(割る)か」「当節の/されちまうから」と解釈されているため。

**表現自体の問題:**

- **テザインみたいのやったり、～**

大変にくだけた文の場合には、「の」の用法自体が、変化している。

しかしながら、重要な助詞の誤りを多く発見することとの兼ね合いにより、これらの過剰検出は望ましくないながらも、許容範囲内であると思われる。

## 5. 今後の課題

助詞の用法は日本語校正の大いなる課題であり、今後とも、形態素解析の向上や、より多くのパターンを発見することにより、再現率・適合率の向上に取り組んでいきたい。

## 参考文献

〔1〕奥村薫："日本語文章校正ツール Chanterelle －入力ミス及び表記揺らぎについて－", 情報処理学会第５５回全国大会 4M-05, (1999)