Emacs系エディタでの日本語の誤り探索[1]

3V－9

川口 湊[2]　蓮井 洋志[3]
福井大学工学部情報工学科[4]

1. はじめに

学術論文のような明晰な日本語表現が要求される文書を執筆するときに利用することを想定した日本語の誤り探索のためのツールを開発した。Unix ワークステーションの環境であれば原則として機種を問わず使用できるが、特にEmacs系エディタで利用する場合に適した設計になっている。(1) 参照する辞書にない自立語を使っていないか、(2) 自立語に続く付属語が妥当であるか、(3) 読点などが欠落していないか、(4) 括弧のバランス、(5) 文体の統一がとれているか、などを辞書を参照しながら探索する。誤り探索の対象となる文書は通常仮名漢字変換プログラムを用いて作成されていることから、形式的には文法規則には合致している誤りが少なくない。これらを検出するために、複合名詞や複合動詞はすべて辞書に見出し語として登録されていることが必要である。正しい文章を誤って誤りと指摘する個所が多くてもよいから、探索すべき誤りを見逃さないことに重点を置く。分野毎の複数の辞書の組み合わせを任意に選択できる。

2. 形態素解析

文章を'字種切り法'により第1近似の文節を求め、形態素解析を行なう。成功しなければ、その近似文節の右端から順に平仮名1文字づつを削っては、この過程を成功するまで再帰的に繰り返す。

2.1. 自立語

平仮名以外の字種の並びはすべて自立語として扱い、漢字列と片仮名列はそれぞれの辞書を検索して品詞情報を得る。(平仮名列の自立語は後で別途処理する。) 記号と英字は名詞、数字は数詞とする。記号、英字、数字の並びは白文字列(使用者が長さを指定できる)または2行以上連続した空行がその直後に続く場合は、誤り探索の対象から外す。(1) 複合名詞、(2)「突然雨が降りだした」という文章の「突然雨」の部分のような副詞+自立語、(3) 接尾語のついた単語、は最長一致法を使って辞書検索をする。

2.2. 付属語

自立語の品詞情報を基に、後に続く平仮名文字列が付属語として妥当であるか検定する。付属語を構成する助詞や助動詞のつながり方をGNU Bisonのための生成ルールとして記述し、平仮名文字列が付属語として受容できるかどうかを判断する。検定で不適となる原因としては、(1) 付属語としては許されない平仮名列、(2) 自立語に続く平仮名文字列のなかに平仮名の自立語が存在する、(3) 語幹に基づいて仮定した自立語が正しくない、の3つの場合がある。

検定に失敗すると、(2) が原因であると想定して、平仮名文字列のなかに (a) 平仮名のみの自立語、または (b) 平仮名で始まる自立語、が存在するか探索する。

これにも失敗すると、(3) が原因とみなして、同じ語幹について次の自立語候補を探す。

2.3. 付属語解析

付属語解析のためのパーサをBisonを用いて作成した。Bisonに与えるパーサの生成規則は、文献[1]に示されている付属語の規則を基本としたが、誤り探索に使うためにこれを次のように変更した。

(1) 名詞に続く付属語の規則を追加した。(2) 助動詞として「みたいだ」、「つもりだ」を追加した。(3) 生成ルールが機械的に「したのであるのである」のような冗長な表現を受容するのを防ぐために同じ意味の助詞や助動詞は二度以上、繰り返すと誤りとする。(4)「したそうだろう」のような、意味の通らない表現を防ぐために、特定の助詞や助動詞の前には特定の意味をもつ助詞や助動詞が来ることを許さない。(5)「見える」、「書ける」などの可能動詞は、それらが辞書に登録されていないときは、「見る」、「書く」が辞書にあればそれから生成できるものとして誤りとはせずに処理するが、警告を出す。

---

[1] Japanese Language Spell Checking with Emacs-type Editors

[2] Minato Kawaguti

[3] Hiroshi Hasui

[4] Fukui University, 9-1, Bunkyo-3, Fukui, 910 Japan

3. 文節間の相関関係

辞書は特殊な単語についての属性情報を含む。また、特定の付属語に対して属性情報を持たせるようにBisonの生成規則を作成した。属性情報は、(1) 文節間の「相関関係」を検査する、(2) 文体に課した制限に適合する表現型式であるかどうかを判定する、という目的に使用する。

文体との整合性の検査の例としては、動詞の「おる」や「したなら」(「いる」、「したならば」ではないか)を検出させる場合などがある。

「相関関係」を次節に列記する。例えば、「こと」「もの」などの形式名詞の前の文節が必ずその単語を修飾していなくてはいけないし、「したことがある。」は正しいが、「したのはことがある。」は誤りである。

漢字の語幹であるときには、辞書の検索は複合動詞辞書から行なう。

3.1. 「相関関係」の種類

1. 動詞の連用形の次は、(a) 複合動詞の後段、(b)「やすい」「にくい」などの特定の単語、(c) 読点、のいずれかが続くが、辞書検索の順位から(a)は除外される。(b)、(c) のいずれにも該当しなければ、'読点が足りない' と警告する。

2. さ行変格活用動詞「関する」「対する」「際する」は、格助詞「に」の後でなければ、誤りとする。

3. その文中に否定形がないといけない副詞「全然」、副助詞「しか」「さえ」、に否定形を伴う文が従わなければ誤りとする。これは文を処理し終わる時にのみチェックする。

4. 型式名詞「こと」「もの」「とき」「ころ」「とおり」「ため」「ほう」「わけ」などは、前の文節が修飾形をしていないと誤りとする。

5. 接続詞「間」「限り」「ないしは」などは前の文節が終止形をしていなければ誤りとする。

4. 辞書

(1) 辞書の見出し語、(2) その語の情報にアクセスするためのオフセット値、(3) 辞書識別子、を構造体要素とするリンクリストとメモリー上のハッシュ表の組み合わせにより効率よく辞書情報を読み出す。

5. 処理例

次のような見出し語数の辞書を用いて次に述べる試験を行なった。

基本辞書— (1) 漢字 : 24462 語、(2) 複合動詞 : 110 語、(3) 片仮名 : 2228 語、(4) 平仮名 : 644 語。

ユーザー定義辞書— (1) 漢字 : 209 語、(2) 複合動詞 : 39 語、(3) 片仮名 : 61 語、(4) 平仮名 : 27 語。

31,786 Byte の TEX ファイル (物理学の論文原稿) に対して行なった誤り探索の所要時間は、Sony NWS-3860 (BSD 4.3) の場合で 1 分 01 秒、DEC 3000-400 AXP (OSF/1) では 33 秒であった。244 個所を誤りと判定したが、そのほとんどは用いた辞書にない専門用語、固有名詞、平仮名の副詞、が繰り返し使われていたためである。この例では Emacs 系のエディタ (Njove) で作業を行なって、2 分程度で 244 個所の候補から 17 個所の真に修正すべき点を検出するのに成功した。

チェック項目の分類	誤り候補	真の誤り
見出し語にない	155	1
付属語が正しくない	59	1
読点が必要	19	9
不要な読点	2	2
文節の相関が不適当	2	0
仮名/語尾の揺らぎ	2	1
不適切な助詞・助動詞	1	0
付属語が長過ぎる	1	0
正統的な書き言葉でない	3	3

6. 文献

[1] 佐伯哲夫 : "陳述—文末の構成", 日本語と日本語教育—文法編—, 文化庁, *pp.*95–117, (1973)

[2] 新川尚子 : "動詞、形容詞、副詞のとりたて", 概説・現代日本語文法, 桜楓社, *pp.*140–142, (1991)

[3] 上田勝彦 : "補助的な単語", *loc. cit.*, *pp.*148–158