5B－3

形態素解析における接続情報による複合語の誤表記の検索[1]

蓮井 洋志[2]　川口 湊[3]　小倉久和[4]
福井大学工学部情報工学科[5]

1 はじめに

複合語の解析にはさまざまな試み[1],[2]がある．確率的構造解析は複合語の分割の曖昧性の解消には有効な方法であるが，複合語の誤表記の検索の方法としては問題があると考える．確率的に正しい複合語が使用例の見たことのない複合語であることがしばしばあると思われる．筆者は推敲支援システムの複合語の誤表記の検索においては，ユーザが正しいと認めたことのない複合語は検出対象にすべきであると考える．その場合，漢字複合語をすべて辞書に登録する[3]方法は，一番正確であるかも知れないが，記憶容量を圧迫し，検索時間をいたずらに長引かせる要因になる．

本論文では，論文推敲支援システムの一機能として，複合語の誤表記を二対相関の接続情報を利用して，検索する方法について述べる．文章の形態素解析を行い，複合語を語基に分割し，複合語になり得る語基の接続関係をユーザの判断で適宜登録する．接続関係が登録されていない複合語は検出対象とする．

2 論文推敲支援システム nspell

論文推敲支援システム nspell は，文章の推敲に必要な情報を文章から抽出することを目標として開発した．誤字・脱字，読点の打ち方の誤り，片仮名の表記の揺れ，一つの付属語連鎖内で付属語の意味関係が冗長である，あるいは矛盾している表現などを検出対象にしている．

2.1 エディタとの連動

福井大学情報工学科で開発されたエディタ Njove あるいは Emacs の文書作成機能の中の一機能として nspell は利用される．nspell はエディタの子プロセスとして起動され，その出力をエディタが利用して推敲を支援する．nspell が出力する問題点のメッセージには，行数と桁数が書かれており，エディタのカーソルはそのポイントに飛ぶようになっている．また，リージョンの文字列を自立語の表記とみなし，その品詞情報を対話的に特定し，辞書に自立語の登録・削除，本論文で述べる接続情報の登録・削除，自立語の検索などを容易にできるようになっている．

2.2 形態素解析

論文推敲支援システム nspell は，形態素解析を行い，その結果をもとに問題点の抽出を行う．nspell の形態素解析は文節単位で行われる．字種境界を利用した総当たり方式を参考にしている．考え得る文節の分割パターンをすべて抽出し、それらを形態素解析結果から得た情報をもとに絞り込む．

文節の分割パターンの生成方法は，自立語を辞書引きし，自立語の品詞をもとに自立語の後のひらがな文字列を付属語パーザで解析する．パーザが解析に成功した場合，そのひらがな文字列は付属語であるとみなす．付属語パーザは，解析するひらがな文字列の前に接続している自立語の品詞にあった付属語であるかどうかを判断し，解析して問題点を検出するパーザである．

2.3 複合語の解析

複合語の解析は自立語の辞書引きで行う．複合語は自立語の語基から構成される．自立語の辞書には品詞情報と単語の表記の対になったものを1エントリとして，64005 エントリを登録した．品詞情報の中には，複合語の語基になり得るか，それとも単独でしか用いないのかなどの複合語に関する情報が入っている．文字列が単語に分割でき，構成する単語が語基になり得るものの集合である時に，その文字列は日本語の複合語になり得ると考える．

2.4 複合語接続情報

接続情報とは，複合語を構成する語基が次の語基に接続することが正しいかどうかの情報のことである．そのため，接続情報は複合語の語基になり得る単語しか持たない．接続情報は接続情報辞書に登録されている．もし，接続情報が辞書に登録されていない場合は，誤表記の複合語として検出する．

(1) 登録・削除

複合語接続情報はエディタ内から容易に登録ができる．検出された複合語が正しければその語の接続情報の登録を行い，誤った複合語であればテキストを訂正

[1]Search Mistaken Compound words, by Connection data between words of morphological analysis.

[2]Hiroshi Hasui

[3]Minato Kawaguti

[4]Hisakazu Ogura

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

する．ファイル全体の複合語が正しければ，ファイルの中のすべての複合語の接続情報を登録する機能も付加した．

接続情報は最初は何も登録されていない状態になっている．ユーザが正しいと思うもののみをユーザが自分の手で登録する．容易に学習できることが本方式の大きな特徴である．

(2) 語基分割の曖昧性

複合語は表記によっては，2通り以上の分割が可能である場合がある．そのような複合語の接続情報の登録・削除では，分割に関する曖昧性を解消せずにすべての分割方法に対して，接続情報を登録・削除する．接続情報のチェックではどれか一つの解釈に対して，接続情報が登録されていれば正しいとする．

(3) 接頭辞・接尾辞，片仮名の扱い

接尾辞は単独では辞書に登録せずに，接尾辞の接続した単語の表記を登録する．複合語の中には，接尾辞が同じでも，接続する語基は異なる単語が多いからである．接頭辞は複合語の語基にしかならない単語として登録する．

また，品詞情報が名詞である片仮名は辞書登録しない．片仮名は人によって表記の揺れが生じ易く，それらの表記をすべて登録するのは無駄であると考えたためである．また，片仮名の大半が複合語の語基になる名詞であり，形態素解析時には片仮名文字列は，複合語の語基となる名詞として処理すれば大きな問題はない．片仮名文字列は接続情報はなくても複合語の誤りとみなさない．片仮名の表記の揺れの検出は，形態素解析を行わずに文献 [4] に書かれていた，片仮名文字列を抽出し，50音順にソートする方法を用いる．

(4) 接続情報辞書のフォーマット

接続情報辞書はバイナリファイルで接続情報は以下のように登録される．

```
'#'(接続される単語自身の辞書 id)(接続される単語自身の単語 id)' '{(接続する単語自身の辞書 id)(接続する単語自身の辞書 id)' '}+\0
```

{}+は中括弧内の文字列の一回以上の繰り返しを表す．() で括られている部分は id を表す数字の文字列ではなく，id を表す数字のコードである．

接続情報を持たない単語は接続情報辞書の中では，以下のように表される．

```
{\n(接続情報を持たない単語の連続する数)}+
```

接続情報を持たない単語の連続する数が256以上の場合に，上のフォーマットが繰り返される．

3 評価実験

同じ分野の論文を3つ学習した後に，それらとは異なった誤表記を5つ含んだ論文に対して複合語接続情報を用いた誤表記の検索機能を使用した結果，20719 byte のファイルを1分34秒で解析ができた．5つの誤りはすべて検出され，検出過誤は146個であった．3つの論文を学習した結果，接続情報辞書に登録されている接続情報は3049個で，接続情報辞書の大きさは21644 byte であった．複合語を辞書にそのまま登録する場合は，単純計算をすると1 entry に約15 byte 必要なので，45735 byte くらいであると思われる．

また，接続情報を用いなかった場合，複合語は462個存在した．316個の学習効果があった．しかし，検出過誤が146個というのはユーザ側からすると多過ぎる．

4 まとめ

本システムは，後述の問題点はあるが，ほとんど以前に使用されたことのある表記の複合語だけを受容する．評価結果を見ると誤った表記を逃していないという点では，非常に有効な方法であった．

接続情報ファイルの大きさはすべてを辞書登録した場合と比較して，約1/2であった．これは，複合語の登録語数を多くすればするほど，必要とする情報量の差は大きくなると思われる．

問題点としては，(1) 検出過誤が多くなること，(2) “システム”+“管理” と “管理”+“主義” を接続情報として登録すると，「システム管理主義」が正しい表記とされること，などが挙げられる．

(1) の問題に関しては，登録しない複合語はすべて検出されるために，新しい事実を書くことを目的とする論文では，学習効果が少ないのは仕方がない．論文を書きながら，そのたびに接続情報の登録を行えば有効な機能になるだろう．

5 *参考文献

[1] 武田浩一，藤崎哲之助:統計的手法による漢字複合語の自動分割，情報処理学会論文誌，Vol.28，No.9

[2] 西野哲朗，藤崎哲之助:漢字複合語の確率的構造解析，情報処理学会論文誌，Vol.29，No.11

[3] 福島俊一，佐々木伸太郎，赤石沢元博，竹元義美:日本語文書構成支援システム St.WORDS，情報処理学会第45回全国大会論文集，3-275 (1992)

[4] 橋本敏彦，藤田憲治，山口哲弘:誤った日本語に気付き始めたワープロ，NIKKEI BYTE，No.1 (1995)