

## 日本語漢字仮名変換における同形語の読み判別

6 C - 7

鈴木 和洋

日本アイ・ビー・エム株式会社 東京基礎研究所

### 1. はじめに

日本語漢字仮名変換において、複数の読みを持つ単語をその文脈に合わせて正しく読むことは、日本語テキスト音声合成や点字翻訳のシステムにおいて、その正読率に大きく寄与する。しかしながら、文脈をとらえることは、談話や意味の構造をとらえる必要があり、深い構文解析や意味解析などの処理が必要となってくる。ここでは、文脈や意味の解析を行わず、特定の形態素の並びの環境をテーブルとして持ち、そのテーブルを用いて効率的に同形語の読み分けを行う方法の検討を行った。なお、漢字列における環境を考慮した読み判別の同様の研究[1]がなされているが、今回は、形態素解析を有するシステムにおいての実現性を考慮して検討をしたものである。

### 2. 従来の方法の概略

日本語の漢字仮名変換において、「行った（いった、おこなった）」「工夫（くふう、こうふ）」などの同じ表記で幾通りかの読みを持つ単語をどのように読み分けるかは、重要な研究課題となっている。KDDの清水らは、同形語の読み分け方法として、単語カテゴリ、品詞、文末タイプ、音訓といった単語の属性を用いた方法を提案している[2]。しかしながら、これらのルールを構築するためには、大量の読み付き文書データによる評価と改良といった試行錯誤を繰り返していく必要があり、ルール蓄積にはかなりの工数を必要とする。

一方、文書校正支援の分野で、同音語の誤った使用をチェックするための方法として、それぞれの単語に対して直前／直後／近隣に出てきやすい語（手がかり語）を列挙した辞書を用いた方法が提案されている[3]。この方法では、単語の意味や文脈などを考慮せず表層のみの解析で同音語の検出がある程度可能になっている。

### 3. 点字文書を利用した読み付きコーパス

上記の同音語の判別は、基本的に同形語についても応用できるが、実際に同形語判別のための辞書を作るためには、大量の読み付き文書データを必要となる。こうした大量の読み付き文書データを作成す

る方法として、電子化された文書と同じく電子化された点字文書との対応を取ることが考えられる。点字は基本的に仮名列の並びとして表現されているため、点字ファイルから読みのファイルは簡単に作ることが可能となる。こうした点字文書からのコーパスを用いることによって、読み判別の精度が大きく改善され、また比較的使用頻度の少ない用例も扱うことができる。

### 4. 形態素列環境テーブルによる同形語判別

この方法では、大量の読み付き文書データ・ベースから同形語が、どのような形態素環境下で、特定の読みを探り得るかをテーブルとして持ち、そのテーブルを用いて、同形語が出現した場合に、そのテーブルを用いて読みを与えるものである。この方法における利点には、以下のものがある。

- ・形態素解析を有するシステムへの移植が容易
- ・修正などのメインテナンスの容易さ

さらに、後述する形態素列環境のルール化によって、品詞、格フレーム、単語シソーラスなどの情報から未知環境への対応が可能である。

#### 4. 1. 形態素列環境テーブルの作成

例えば、2つの読みy1, y2を持つ語Aを考える。Aは下記のようないくつかの形態素並びにおいてそれぞれy1かy2のどちらかの読みを探る。

...X<sub>n-2</sub>, X<sub>n-1</sub>, X<sub>n</sub>, A(y1), X<sub>n+1</sub>, X<sub>n+2</sub>...  
...Y<sub>n-2</sub>, Y<sub>n-1</sub>, Y<sub>n</sub>, A(y2), Y<sub>n+1</sub>, Y<sub>n+2</sub>...

このとき、y1の読みを探るA(y1)の形態素環境として、[...X<sub>n-2</sub>, X<sub>n-1</sub>, X<sub>n</sub>, X<sub>n+1</sub>, X<sub>n+2</sub>...]を、y2の読みを探るA(y2)の形態素環境として、[...Y<sub>n-2</sub>, Y<sub>n-1</sub>, Y<sub>n</sub>, Y<sub>n+1</sub>, Y<sub>n+2</sub>...]を各環境テーブルのデータとして、文書データデータ・ベースにおける出現頻度とともに保存される。さらにA(y1), A(y2)の各環境テーブルで競合するデータについては、出現頻度の少ない方のデータを削除する。

#### 4. 2. 形態素列環境のルール化

形態素環境を、品詞、格フレーム情報や単語シソーラスを用いてルール化し、環境データの削減や未知環境への対応を行うことができる。たとえば、

「行つ（おこな）」については、文法情報を用いて、次のようなルール化が可能である。

「(サ変名詞)+を+行つ」

「行つ+（連体語尾）+（サ変名詞）」

また、単語シソーラス情報を用いて、以下のようなルール化も可能である。

「（時を表す名詞）+に+行つ」

#### 4. 3. 形態素列環境データを用いた読み判別

読み判別では、対象となる同形語Aに対してその環境が、A(y1)の環境と類似するか、A(y2)の環境と類似するかを判別関数で計算する。判別関数では、それぞれの環境を示す形態素の有無、およびルールとの適合度などが評価される。

#### 5. 実験と評価

前述のシステムの妥当性を評価するために「行つ（いつ、おこなつ）」「通つ（とおつ、かよつ）」の読み分けを検討した。今回は、形態素解析を行わず、形態素列環境の抽出として、ターゲットとなる同形語の前に存在する平仮名列（付属語列に相当）、および漢字／片仮名列（自立語に相当）を環境データとして拾いだして環境テーブルを作成した。判別の際には、抽出のときと同様に、同形語の前の平仮名列、漢字／片仮名列とともに、同形語の直後の漢字／片仮名列のマッチングを行った。漢字／片仮名列のマッチングは、比較するデータの一部分が環境データと一致するものも対象とした。

判定の手順は以下の通り。

(1) まず、漢字／片仮名列として、マッチングするものがあるかどうかを比較する。このとき、両方のデータ・ベースにマッチングするものがあれば、出現頻度の高い方を選択する。(2) 漢字／片仮名列のデータ・テーブルにマッチングするものがないか、あるいは両方にマッチングものがあるが出現頻度が同じ場合、平仮名列のデータ・テーブルにマッチングするものがあるかどうかを調べる。このとき、両方にマッチングするデータがある場合は、漢字／片仮名列のときと同様に、出現頻度の高い方を選択する。

環境データの作成に使用した読み付きデータは、EDRの読み付きデータと日本経済新聞から得られたサンプルの半分を使用した。評価はクローズのデータとして環境データの作成で用いたサンプルを、オープンのデータとして残り半分のサンプルを用いた。また、今回は、形態素解析を行っていないため形態素列環境のルール化およびその適用は行わなかった。

表1、表2に、「行つ」および「通つ」の判別誤りの結果を示す。

表1. 「行つ」における判別誤り

	クローズ (2500サンプル)	オープン (2561サンプル)
オコナツ(タ)	50	132
イツ(タ)	72	225
計	122 (4.88%)	357 (13.96%)

表2. 「通つ」における判別誤り

	クローズ (500サンプル)	オープン (529サンプル)
トオツ(タ)	6	33
カヨツ(タ)	7	56
計	13 (2.6%)	89 (16.82%)

オープン実験では、やはり予想されるようにデータが十分でないためか誤り率が高い。判定手順として、自立語部と付属語部を対等に評価することで誤り率は減少すると考えられる。また、さらにデータを多くし、形態素列環境のルール化を行うことにより、精度は向上すると考える。

#### 6. おわりに

今後は、点字文書を使用した読み付きデータを作成し、大量の文書すべての同形語について、形態素環境テーブルを作成するとともに、品詞、格フレーム、単語シソーラスを用いた形態素環境テーブル上のルールの自動生成などを進め、テキスト音声合成や点字翻訳システムに組み込む予定である。

#### 文献

- [1] 鳥原信一：「漢字N-gramを用いた読み付与システム」，情報処理学会第54回全国大会講演論文集，(1997)。
- [2] 清水徹、樋口宣男、河井恒、山本誠一：「隣接単語間の結合関係に着目したテキスト音声変換用形態素解析処理」，音響学会誌 51, pp. 3-13 (1995)。
- [3] 奥村薫、脇田早紀子、金子宏：「日本語校正支援における同音語誤り検出－警告レベルの提案」，情報処理学会第49回全国大会講演論文集，(1994)。