単語の共起の関係と用例を用いた同形語の判別

池田裕和　　宮崎正弘

新潟大学大学院自然科学研究科

1　はじめに

漢字仮名混じりの日本語文をコンピュータで解析する際に、表記が同じで読みが異なる「同形語」による品詞や読みの誤りが問題となる。語の読みや品詞が異なれば語の意味も異なる場合があり、この同形語によって本来とは違った意味の文が解析結果として出力され、日本語文解析の精度低下を招く。

本稿では解析精度向上のため、同形語の品詞や読みを一意に判別することを目的とし、同形語について名詞句、複合名詞内の単語の共起関係、用言の格パターンや読みの使用頻度、用例などの単文内の情報などを利用して、同形語を自動処理で一意に読み分ける方法を提案し、その有効性を確認した。

2　同形語の種類

同形語には大別して名詞類の同形語と動詞類の同形語の２種類がある。ここでは同形語の種類を具体例とともに述べる。

- 名詞類の同形語
  - 複合名詞内係り受けによる判別　→「関東**平野** (へいや/ひらの)」etc.
  - 格助詞「の」を介した名詞の判別　→「木の**根** (ね)」「方程式の**根** (こん)」 etc.
  - 名詞単体での判別　→　「**工夫** (こうふ) が**工夫** (くふう) する」 etc.
- 用言類の同形語
  - 助詞、助動詞の第一音節の清濁による判別　→　「**研** (と) **い** だ」「**研** (みが) **い** た」 etc.
  - 用言の格要素による判別　→　「評価を**行** (おこな)**った**」「野を**行** (い)**った**」 etc.

ところで、「複合名詞内係り受けによる判別」は複合語の構造解析部で同形語の判別をも行うことができる。また、「名詞単体での判別」はどのような用言の格要素であるかによって判別できる場合と、一文内では判別できず、全体の文脈の中で判別が行われるケースもある。

本稿では「格助詞『の』を介した名詞の判別」「助詞、助動詞の第一音節の清濁による判別」「用言の格要素による判別」、「数詞関連の同形語」についてその判別法を述べる。

3　同形語自動判別法

本節では、コンピュータ上で同形語自動判別がどのように行われるのかを述べる。自動判別の位置づけを、図１に示す。

まず入力された文は形態素解析によって形態素に分割される。なお、本稿では漢字仮名混じり文を対象とする日本語形態素解析システム Maja[1] を使用した。

次に、形態素に分割された文を形態素解析後処理 [2] に渡す。同形語自動判別はこの形態素解析後処理内の一処理であり、入力された単語列を単語書き換えルールファイルと照合することで行う。単語列がルールに合致していれば、単語列を書き換え正しい読みをあてる、という手法で同形語自動判別を行う。

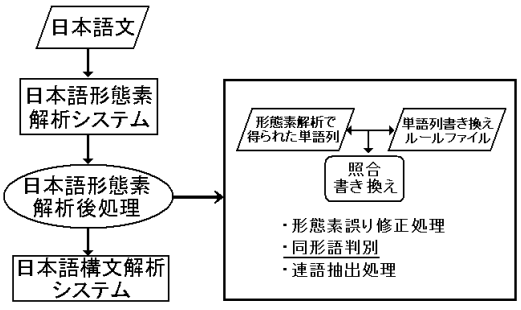

図 1: 同形語自動判別の位置付け

4　格助詞「の」を介して結合された名詞の判別

名詞が格助詞「の」を介して他の名詞を修飾する場合、判別の対象となる名詞に当該名詞の前方、もしくは後方に格助詞「の」を介して接続できる名詞の階層的意味属性を与えておき、判別の対象となる名詞の前方、または後方に格助詞「の」を介して接続した名詞の意味属性と一致する、もしくは妥当である属性を持った読みを選択する。また、一般的な慣用表現をあらかじめ辞書に収録し、自動判別をすることも可能である。具体例を、図 2 に示す。

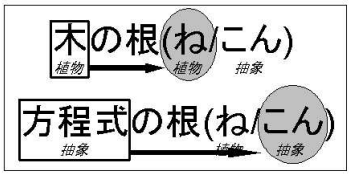

図 2: 具体例

5　助詞、助動詞の第一音節の清濁による判別

5.1　判別方法

これは、動詞の音便形に助動詞 (「た」「だ」「て」「で」)、助詞 (「ても」「でも」) が直後に接続されている場合、その助詞、助動詞の第一音節が清音か濁音かによって読みを判別する。なお、この方法で判別が可能な表記と読みの一例を表 1 に示す。

表 1: 判別可能な同形語の一例 (清濁)

表記	読み (後続が濁音)	読み (後続が清音)
研い	とい	みがい
急い	いそい	せい
傾い	かしい	かたむい
塞い	ふさい	せい

5.2　書き換えルールファイルの実装

動詞の音便形、助動詞の字面、品詞情報を元に書き換えルールを以下のようにコード化し、書き換えルールファイルを実装した。

```
## 研い (みがい) た (既定の助動詞 終止形)
研　　21[12]0
い　　21[12]4
た　　c316
=
研　　2110
い　　2114
た　　c316

…
```

6　用言の格要素による判別

用言の前方には用言の格要素となる名詞文節がある。そこで、判別対象となる用言の前方にあり、当該用言を修飾する名詞文節の中から格定詞 (副助詞) を手がかりに、用言の必須格となりうる文節を見つけ、当該文節の名詞がもつ意味属性が用言の格パターンで指定された意味属性と一致する用言を選択する。なお、用言の格パターンにマッチしない場合は、デフォルト読みとして読みの頻度に基づく優先読みを与える。この方法で判別が可能な表記と読みの一例を表 2 に示す。

表 2: 判別可能な同形語の一例 (用言の格要素)

表記	読み (優先度が高)	読み (優先度が低)
行った	いった	おこなった
断った	たった	ことわった
終った	おわった	しまった
通った	とおった	かよった
勝った	かった	まさった

6.1　判別優先度フラグを用いた優先読みセット処理

Maja 辞書に収録されてある単語優先度フラグを利用して優先度の高い読みにあらかじめセットする処理を

自動で行う。なおこの処理により、判別の対象となる文がルールに合致していなくとも、比較的使用頻度の高い読みが得られるので解析精度の向上が期待できる。

6.2 名詞の意味属性リスト 付与作業

Majaによる形態素解析の後、文中に現れた名詞に対して辞書引きを行い、名詞の階層的意味属性 [3] を抽出する。この意味属性は木構造で体系化されており、根から当該単語の意味属性までの経路をリスト化して単語に「意味情報」として付加した。

6.3 変換ルールファイルの実装

同形語や助詞の字面や品詞、そして名詞の意味属性の情報を元に以下のようにコード化し、ルールファイルを実装した。

```
## 動作名詞 (人間活動/変動) に * 行 (い)った。
*       12\w0,.*(1236|2064)
に      d130
::      ----
行      21[9a]0
っ      21[9a]4
=
*       12*0
に      d130
::      ----
行      21a0,---,"い"
っ      21a4
…
```

6.4 数詞関連の同形語

数詞関連表現にも、「一足 (いっそく/ひとあし)」のように、同形語が存在する。また、語によっては「一分 (いっぷん/いちぶ/いちぶん)」のように、3 通り以上の読み方をする語や、「九九 (きゅうじゅうきゅう/くく)」、「一 (いち/はじめ)」のように、数詞単体で一般語や固有名詞と同形となりうる語も存在する。

複合語内に現れる複合語は複合語構造解析部で対処できるものもあるが、本稿では用言の格要素による同形語の判別と同様に、単語の共起関係や用例を用いて判別することとした。

6.5 評価実験結果

用例辞典 [4][5] を参照して、同形語を使った試験文 (258 文) を作成し、同形語自動判別が正しく行われているかを検証した。結果を各同形語別に表 3 に示す。

この結果より、まず読みの頻度情報を用いない場合よりも用いた場合のほうが高い正解率をあげた。また、各同形語別の正解率をみると「行 (い/おこな)った」、「通 (とお/かよ)った」などのように読み方次第で意味が違い、読み分けが必要なもので比較的高い正解率をあげた。逆に、正解率の低かったものは「計 (はか/はから)った」、「語 (かた/かたら)った」のような読み方で意味に大きく差のない同形語であった。

数詞関連に関しては、一日 (いちにち/ついたち/いちじつ)、一味 (いちみ/ひとあじ)、一足 (いっそく/ひとあし)、一分 (いっぷん/いちぶ/いちぶん) の四つについて試験文を作成し、同形語判別実験を行った。頻度情報を用いない場合はどちらか一方の読みのみを採用する傾向が強く、正解率は半分以下という、精度の低い結果となった。頻度情報を用いることで、正解率が向上した。しかし、一足 (いっそく/ひとあし) に関しては、正解率がそれほど向上しなかった。

7 おわりに

本稿では、Majaによる形態素解析の後、書き換えルールファイルと入力文との照合により同形語の自動判別を行う手法を提案し、用言の格要素による同形語や数詞関連の同形語について判別実験を行った。その結果、平均約 85 %という正解率を得た。

しかしながら、同形語ごとの正解率にばらつきがあった。そこで正解率の差を埋め、なおかつ判別精度の向上のために書き換えルールファイルの拡張が必要である。また本稿では、名詞の意味属性抽出の部分で従来型のシソーラスを使用したが、今後は現在試作中の連想型多次元シソーラス [6][7] を利用した同形語判別を試みる予定である。連想型多次元シソーラスによる判別法は、4 節で述べた「格助詞「の」を介して結合された名詞の判別」に特に有効である。また、本稿は試

表 3: 各同形語に対する解析結果

語表記	試験文数	正解数 (読み頻度無)	正解数 (読み頻度有)	正解率 (%)(読み頻度有)
行った	36	32	33	91.7
断った	24	23	23	95.8
終った	24	19	23	95.8
通った	40	35	35	87.5
勝った	25	19	18	72.0
競った	6	5	5	83.3
計った	11	7	7	63.6
語った	7	2	5	63.6
集った	10	7	8	80.0
賜った	5	4	4	80.0
強い	17	17	17	100.0
正しく	15	13	13	86.7
一日	6	2	5	83.3
一味	10	4	9	90.0
一足	11	5	6	54.5
一分	11	4	8	72.7
合計	258	198	219	84.88

験文に単文を用いたが、単文内の情報だけでなく、文脈情報を利用した同形語判別も試みる予定である。

そして、以上の方法を組み合わせ、自然言語処理の応用分野である音声出力や機械翻訳へ適用し、より精度の高い同形語判別を実現したい。

参考文献

[1] 尾嶋, 宮崎:日本語形態素解析システムにおける部分的再試行機構の導入とその効果、情報処理学会第 58 回全国大会,1E-4(1999)

[2] 佐々木:単語連鎖列書き換え規則を用いた日本語形態素解析後処理 (2004)

[3] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林:日本語語彙大系, 岩波書店 (1997)

[4] 小泉, 船城, 本田, 仁田, 塚本:日本語基本動詞用法辞典, 大修館書店 (1989)

[5] 林 他:現代国語用例辞典, 教育社 (1992)

[6] 森田, 宮崎:連想型多次元シソーラスとその意味解析の適用性, 言語処理学会第 12 回年次大会,A4-2(2006)

[7] 小林, 宮崎:既存のシソーラスを利用した仮想シソーラスの構築, 言語処理学会第 13 回年次大会,B5-5(2007)