意味的共起関係を用いた動詞と名詞の同音異義語の仮名漢字変換

元永 靖和　　池原 悟　　村上 仁一
鳥取大学　工学部
{motonaga, ikehara, murakami}@ike.tottori-u.ac.jp

1 はじめに

音声認識やワープロなどで使用される仮名漢字変換には、同音異義語の誤変換という問題がある。この問題に対して、従来、単語の頻度や、最も最近に使用した単語を優先させる方法が使われていた。しかしながら、これらの方法は、単語や品詞間の文法的、意味的な関係を考慮していないため、不自然な変換結果を出力する。

これに対して、近年では、単語間の意味的な関係を考慮した方法により変換精度を向上させる試みがなされている。主な方法としては、(a) 単文内で格関係を持つ名詞と動詞間の単語共起情報を用いた方法 [1]、(b) 大規模な連語共起情報を意味素で記述し、その関係を調べる方法 [2]、(c) 単文内で用言の格フレームを用い、意味的整合性から変換する方法 [3, 4] などが挙げられる。

しかし、これらの方法には以下のような問題がある。

(a) の方法では、単語共起情報のない単語に対しては効果がない。単語共起情報は表記そのものを記述するため、自然言語の持つ量的な性質を考えると、膨大なものとなり収集が困難である。また、単語共起情報の網羅性についても考慮する必要がある。(b) や (c) の方法では、意味素体系を適切に設定し、さらに (c) の方法では、一貫性を持った格フレームを大規模に構築する必要がある。

これに対して、本論文では、結合価パターンを用いた同音異義語の選択について述べる。結合価パターンは、名詞と用言の関係を記述したものであり、文法的、意味的な情報を含んでいる。結合価パターンと類似した格フレームを用いた前述の方法 [3, 4] は、十分な効果が得られていない。また、文献 [3] では、変換の誤りが約 20%減少するにとどまっている。これは、名詞の分類や格フレームの精度が不十分であったためと考えられる。これと比較して、本論文で用いる結合価パターンは、規模が大きく名詞の分類精度が高い。そこで、同音異義語の選択に対する結合価パターンの有効性を調べる。

2 結合価パターンによる同音異義語の選択

2.1 結合価パターンと一般名詞意味属性

結合価パターンは、用言と格要素（名詞＋助詞）の意味的関係を記述したものなので、意味的な制約が生じる。この制約を利用すれば、仮名漢字変換における同音異義語の選択にも応用できると考えられる。

本論文では、日本語語彙大系 [5] に掲載されている「構文意味辞書」の結合価パターンを使用する。これは、日英機械翻訳システム ALT-J/E[1] の日本語解析で発生する意味上の多義を解消するという目的で開発され、日本語の用言を中心とする文型を結合価パターンにまとめたものである。

具体的には、一般文型と慣用表現文型をあわせて約 15,000 件の日英文型パターン対が収録されている（本論文では、日本語の文型パターンのみを用いる）。結合価パターンの例を表 1 に示す。

この結合価パターンでは、格要素の名詞に当たる部分が一般名詞意味属性 [5] で記述されている。一般名詞意味属性とは、名詞を、単語の見方、捉え方に着目して、名詞の意味的用法を整理、体系化したシソーラスである。本論文で使用する一般名詞意味属性体系では、約 40 万語の名詞を、最大 12 段の木構造を構成する 2,710 の属性に分類している。一般名詞意味属性体系の一部を木構造で表した図と各属性に属している名詞の例を示す（図 1）。

[1] NTT が研究開発した日英機械翻訳システム

また、この一般名詞意味属性体系は、木構造を基本構成としているため、上位の意味属性の性質を、下位の意味属性の性質に伝搬・継承できるという性質がある。この上位下位の関係を利用して結合価パターンの照合を行う。

表 1: 結合価パターンの例（かっこ内は一般名詞意味属性名）

（人）	が	（食料）	を	断つ
（人）	が	（場所）	を	発つ
（天体）・（煙）	が	（空）	に	昇る

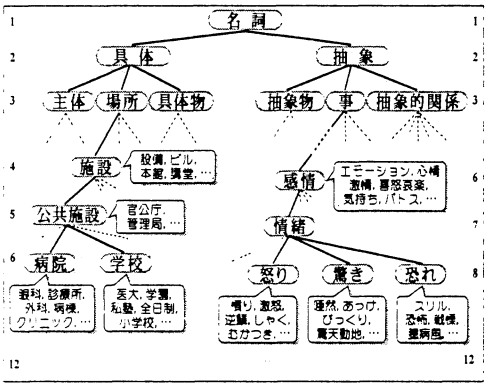

図 1: 一般名詞意味属性体系（一部）

2.2 同音異義語の選択方法

一般に、動詞と名詞には同音異義語が多いため、これらの変換誤りは仮名漢字変換において重要な問題となる。結合価パターンは、動詞と名詞の意味的関係を記述したものであるから、動詞と名詞の同音異義語の選択に有効であると思われる。そこで本論文では、動詞と名詞の同音異義語のみを選択対象とし、その他の単語は変換済みとする。

結合価パターンを用いた同音異義語の選択する手順を以下に示す（図 2）

1) 入力文から、着目する動詞とそれに対応する格要素を取り出す（図 2 の例では①）。
2) 着目する動詞の「読み」に対応する結合価パターンを読み出す（図 2 の②）。
3) 入力文側の格要素の名詞の意味属性を調べ（図 2 の③）、結合価パターン側の格要素の意味属性と照合する。
4) 一般名詞意味属性体系上で、入力文側の意味属性が結合価パターン側の意味属性の配下にあれば、その格要素は条件を満足したと見なす（図 2 の④）。
5) 手順 4) を残りの他の格要素全てに対して行い、最も条件を満足した結合価パターンを元に、同音異義語の選択を行う。図 2 の例では、④の表から「断つ」の結合価パターンが選択され、それぞれ「こうぶつ」は「好物」が、「たつ」は「断つ」が選ばれる。

従って、同音異義語が動詞ならば、最も条件を満足した結合価パターンに記述されている動詞が選択され、名詞であれば、最も条件を満足した結合価パターンに対応する名詞が選択される。

3 実験と考察

結合価パターンを用いた同音異義語の選択の有効性を評価するために、計算機で実験を行った。以下に実験方法を示す。

3.1 試験文

同音異義語の選択の対象となる試験文は、以下に示す条件で収集した（表 2）。なお、同音異義語は試験文 1 文に対して 1 語ずつ含んでいる。

動詞試験文：動詞の同音異義語を選択する実験

情報処理振興事業協会が作成した計算機用日本語基本動詞辞書 IPAL（以下、IPAL 動詞辞書）に収録されている基本動詞（861 語）を、ALT-J/E 単語辞書に収録されている動詞（約 6.000 語）と比較した（終止形のみ）。この中で同音異義語の関係をとるのは、よみで 159 種、漢字表記で 388 語存在した。この IPAL 動詞辞書の各動詞に付随して収録されている用例文を、単文に直し、動詞の試験文（500 文）とした。

名詞試験文：名詞の同音異義語を選択する実験

計算機用日本語基本名詞辞書 IPAL（以下、IPAL 名詞辞書）に収録されている基本名詞（1.081 語）を、ALT-J/E 単語辞書に収録されている名詞（約 40 万語）と比較した。この中で同音異義語の関係をとるのは、よみで 500 種、漢字表記で 1.975 語存在した。この IPAL 名詞辞書の各

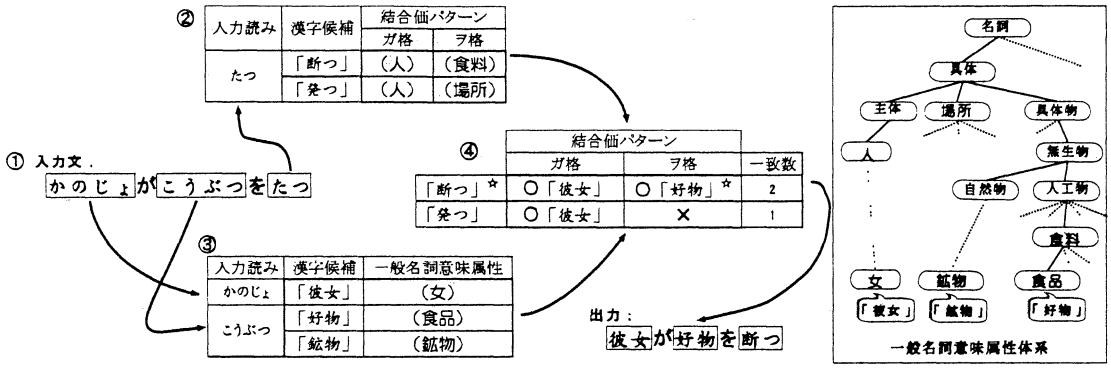

図 2: 結合価パターンによる同音異義語の選択例（かっこ内は一般名詞意味属性の属性名）

表 2: 試験文

	動詞	名詞
IPAL 単語数	861 語	1,081 語
ALT 単語数	約 6.000 語	約 40 万語
同音異義語（よみ）	388 語（158 種）	1,975 語（500 種）
試験文数	500 文	500 文
試験文の例 （《》内は同音異義語）	映画をスクリーンに うつす。 《写す,映す. 移す》 ブッダが難民を貧窮から すくう。 《掬う,救う, 巣くう》 怒りが おさまる 《治まる, 修まる,収まる, 納まる》	たけ でおもちゃを作る。 《丈, 他家,竹》 膝から した が露出する。 《舌,下》 ケガ人を病院に運ぶ。 《病因,病院》

名詞に付随して収録されている用例文を、単文に直し、名詞の試験文（500 文）とした。

なお、正解の表記は、各 IPAL 辞書の表記に従う。IPAL 辞書の表記が複数存在する場合は、いずれか一語以上一致していれば正解とする。

3.2 実験方法

結合価パターンによる同音異義語の選択と比較するため、従来の単語共起情報を用いた方法による実験を行う。実験は以下の 3 つの方法で行い、前述した動詞と名詞の試験文各々に対して計算機で同音異義語を選択させた。

(a) 結合価パターンのみを用いた方法
(b) 動詞と名詞の単語共起情報のみを用いた方法
(c) 結合価パターンと単語共起情報を組み合わせた方法

(b) で用いる単語共起情報は、動詞 1 語に対して名詞 1 語の組を収集し、組ごとに頻度を付与したものを使用した（図 3）。具体的には ALT-JAWS[2] によって形態素解析処理済みの新聞一年分[3]（約 150 万文）のテキストから約 70 万組の単語共起情報の標本を収集した。この単語共起情報を元に同音異義語を選択させる。

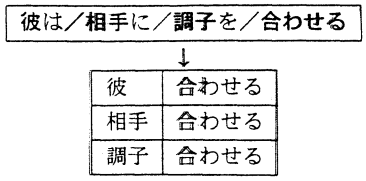

図 3: 単語共起情報の例

また、(c) は、試験文に対して、まず先に結合価パターンを適用させる。その結果、

[2]NTT が開発した形態素解析処理システム
[3]CD-毎日新聞'95 データ集

- 得られた候補の数が適用前と変わらない場合
- 得られた候補の数が 0 の場合

のいずれかの場合に、単語共起情報による評価を優先させるという方法で実験を行った。

3.3 実験の結果と考察

同音異義語を一意に選択できた試験文の割合と、選択対象とする同音異義語の平均多義数の変化値を以下に示す（表 3）。また、実験 (a) の出力結果を表 4 に示す。

表 3: 一意に選択できた試験文（かっこ内の数値は平均多義数の変化）

	動詞	名詞
(a) 結合価パターン	77%(2.9 → 1.4)	26%(3.8 → 2.9)
(b) 単語共起情報	28%(2.9 → 2.1)	37%(3.8 → 2.8)
(c) 両方の組合わせ	79%(2.9 → 1.3)	42%(3.8 → 2.3)

実験の結果、動詞の場合において、正解率が結合価パターンを用いた方法は単語共起情報を用いた方法よりも高くなった。しかし、名詞の場合には逆の結果となった。これは、単語共起情報の収集元である新聞の文と、試験文の収集元である IPAL 辞書の用例文の分野の違いではないかと考えている。

また、(b) の方法では単語共起情報に一致しない試験文が多くあった。これは単語共起情報の不足と考えられる。

(c) の方法では、効果の向上は余り多くなかった。従って、組み合わせの方法を考え直す必要があると考えられる。

4 おわりに

本論文では、動詞と名詞の意味的関係に着目し、結合価パターンを用いた同音異義語の仮名漢字変換の効果を評価した。具体的には、日本語語彙大系に収録されている結合価パターンと単語共起情報を使用し、同音異義語の動詞を含む単文と、同音異義語の名詞を含む単文それぞれに対して実験を行った。この結果、同音異義語の中から正解の表記を決定できた文は、結合価パターンを用いた方法では、動詞の場合で 77%、名詞の場合では 26%、単語共起情報を用いた方法では、それぞれ 28%、37%、これらを組み合わせた方法では、それぞれ 79%、42% であった。これらの結果から、結合価パターンを用いた同音異義語の選択において、特に同音異義語の動詞に対して効果があることが分かった。

表 4: 実験 (a) の出力結果（かっこ内は出力単語）

	動詞試験文	同音異義語数
正解文	彼は花びんに花を さす	5（挿す）
	彼は足が つる	2（攣る）
	彼女から返事が かえる	8（返る）

	動詞試験文	同音異義語数
失敗文	君が玄関の前に たつ	5（発つ）
	彼は元の職場に おさまる	4（修まる）
	彼は愛に うえる	2（植える）

	名詞試験文	同音異義語数
正解文	ここに けんちょう を建設する	3（県庁）
	南西の ほうがく を目指す	3（方角）
	委員長の にんき を尋ねる	2（任期）

	名詞試験文	同音異義語数
失敗文	部外者が こうこう に入る	19（後攻）他
	筆の ほ を揃える	5（歩）
	はら を見せる	3（原）

今後は、単語共起情報の量を増加させた実験と、動詞と名詞の両方が同音異義語の文を対象にした実験を行う予定である。

参考文献

[1] 高橋, 吉村, 首藤 (1996)：単文内での共起情報を用いた同音語処理, 情報処理学会論文誌,Vol.36,No.6,pp.998-1006

[2] 本間, 山階, 小橋 (1986)：連語解析を用いたべた書きかな漢字変換, 情報処理学会論文誌,Vol.27,No.11,pp.1062-1067

[3] 大島, 阿部, 湯浦, 武市 (1986)：格文法によるかな漢字変換の多義解消, 情報処理学会論文誌,Vol.27,No.7,pp.679-687

[4] 牧野, 木澤 (1981)：べた書き文の仮名漢字変換システムとその同音語処理, 情報処理学会論文誌,Vol.22,No.1,pp.59-67

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