自然言語処理　132－10
（1999. 7. 23）

動詞と名詞の意味的共起関係を用いた同音異義語のかな漢字変換

元永 靖和　　池原 悟　　村上 仁一

鳥取大学　工学部

E-mail:{motonaga, ikehara, murakami}@ike.tottori-u.ac.jp

あらまし　本論文では、かな漢字変換における同音異義語の変換誤りの問題に対して、名詞と動詞の意味的関係を記述した結合価パターンを適用した効果について報告する。具体的には、動詞と名詞の同音異義語を含む試験文（4632 文と 504 文）に対して、日本語語彙大系の構文意味辞書に収録されている結合価パターンを適用し、同音異義語を選択する実験を行った。その結果、正解表記を一意に決められた単語が、動詞では 33%、名詞では 22%あり、候補を絞り込めたものを含めると、49%, 37%であった。多義平均数は、動詞では 4.3 から 2.9 へ減少し、名詞では 4.2 から 3.3 へ減少した。また、絞り込めなかった候補をランダムに選択すると仮定し、全体をランダムに選択する場合と比較すると、正解率が動詞では 25%から 48%に、名詞では 34%から 47%に向上した。

キーワード　かな漢字変換、同音異義語、結合価パターン、意味属性

Study of Kana-to-Kanji Conversion using semantic collocations between verb and noun

Yasukazu Motonaga　　Satoru Ikehara　　Jinichi Murakami

Faculty of Engineering, Tottori University

Abstract　We describe effects of valency patterns using semantic collocations between verb and noun for problems of homophonous in Kana-to-Kanji conversion. For 4,632 sentences with homophonous of verb (verb tasts) and 504 sentences with homophonous of noun (noun tests), respectively, we tested selections of homophonous by valency patterns from "Syntactic and Semantic Dictionary" of the "Nihongo Goi Taikei".

As a result, 33% of verb tests and 22% of noun tests could be selected on homophonous, reduced candidates are 49% and 37%, respectively. Average of polysemy is reduced from 4.3 to 2.9 and from 4.2 to 3.3. Accuracy is improved from 25% to 48% and from 34% to 47%, compared with random selections of homophonous.

key words　Kana-to-Kanji conversion, homophonous, valency pattern, semantic attribute

1. はじめに

日本語のワードプロセッサや音声認識に用いられるかな漢字変換は、近年、コスト最小法などの高度な処理や大規模な辞書の使用が可能になったことにより、精度が向上している。しかし、同音異義語に対する変換精度の向上は、かな漢字変換において、依然として重要な問題の一つである。

この問題に対して、従来、単語の頻度を利用した方法や、最も最近に使用した単語を優先させる方法が、一般的に使われていた。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。これに関連した方法として、(a)単文内で格関係を持つ名詞と動詞間の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]などがある。

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、共起情報のない単語に対しては効果がない。さらに、この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、その共起情報をどの程度収集すればよいのかという網羅性も考える必要がある。(b)や(c)の方法では、意味素体系を適切に設定する必要がある。さらに(c)の方法は、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。

本論文では、結合価パターンを用いたかな漢字変換の有効性を調べる。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方法[3,4]は、あまり効果がなく、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。本論文で用いる結合価パターンは、最近、岩波書店より出版された日本語語彙大系[5]に収録されている結合価パターンである。これは、従来より規模が比較的大きく、名詞の分類精度が高い。この結合価パターンを使用して、動詞と名詞の同音異義語の選択の有効性を調べる。

以下、2章では、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、3章では、実験内容とその結果を示し、実験結果について考察する。

2. 結合価パターンによる同音異義語の選択

2.1. 結合価パターンと単語意味属性

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である。これにより用言と名詞との間に意味的な制約が生まれ、この制約を利用して、かな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、日英機械翻訳において、日本語解析で発生する意味上の多義の解消を目的として開発され、日本語用言を中心とする文型を格フレームと類似の結合価パターンにまとめたものである。一般の文型と慣用的表現の文型をあわせて約16,000件の日本語文型パターンにまとめられている[5]。表1に結合価パターンの例を示す。

表1　結合価パターンの例（かっこ内は意味属性名）

用言	結合価パターン
登る	(天体、煙)が(宇宙)に昇る
飲む	(波、河川、海)が(具体)を飲む
収穫する	(人)が(植物)を収穫する
傾げる	(人、動物)が(頭部)を傾げる
駆け出す	(人、動物)が駆け出す
確定する	(名詞)が(名詞)と確定する
生まれる	(感覚、感情)が(心)に生まれる
生まれる	(商品)が生まれる

この結合価パターンでは、格要素の名詞の代わりに一般名詞意味属性[5]で記述されている。

一般名詞意味属性とは、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（図1）は、約40万語の名詞を、最大12段の木構造を構成する2,710の意味属性に分類されている。

また、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[5]という性質がある。この性質により、下位の意味属性を定義するための記述量を減らしている。結合価パターンは、この上位下位の関係を利用してパターンの照合を行う。

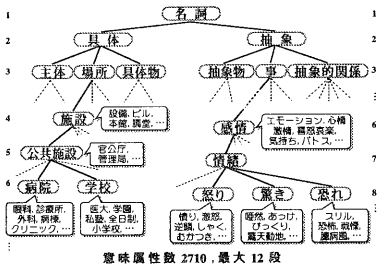

図1　一般名詞意味属性体系（一部）

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重要な問題である。また、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はないかと思われる。そこで、本論文では、動詞と名詞の同音異義語を選択対象とする。

始めに、用言の読みをキーとして辞書から結合価パターンを読み出す。次に、入力文側の助詞をキーとして、読み出した結合価パターン側との対応関係を照合する。対応関係のあった入力文側の格要素の名詞の意味属性を調べ、結合価パターン側の意味属性と照合する。一般名詞意味属性体系上で、結合価パターン側の意味属性が、入力文側の意味属性の上位に位置すれば、その格要素は一致したと見なす。これを他の格要素全てに対して行い、最も一致した結合価パターンを元に、同音異義語の選択を行う。つまり、同音異義語が動詞ならば、その結合価パターンに記述されている動詞が選択し、名詞であれば、その結合価パターンに一致した名詞を選択する。

3. 実験

結合価パターンを用いた同音異義語の選択の精度を評価するために、実験を行った。

3.1. 実験の条件

（1） 前提条件

本論文において以下の用語を定義する。

- 同音異義語とは、品詞、読みとも同じで意味が異なる2つ以上の語と定義する。
- 候補数とは、ある一種類の同音異義語の総数と定義する。
- 選択された候補のことを「選択候補」とし、その数を「選択候補数」と定義する。

また、本論文では、以下の条件を仮定する。

- 試験文の正確な形態素解析、構文解析は得られていると仮定する。
- 実験の対象とする同音異義語のみが「かな」で、そのほかの単語は、正確な漢字へ変換されていると仮定する。

（2） 実験の種類

実験は、以下の2つに分けて行った。

実験1：動詞の同音異義語を選択する実験

ALT-J/E[1]の日本語単語辞書（以下、ALT単語辞書）から同音異義語となる動詞（14,403語5,682種）を抽出し、小説[2]及び新聞一年分[3]（約210万文）に含まれる比較的頻度の高い64語18種の同音異義語となる動詞を選んだ（表2）。こ

[1] 日本電信電話株式会社(NTT)が開発した日英機械翻訳システム
[2] CD-ROM版 新潮文庫の100冊
[3] CD-毎日新聞'95データ集

れらの動詞を含む文を、前述の小説及び新聞から抽出し、単文に直し整理したものを動詞用の試験文（4,632 文）とした（表 3)。図 2 に試験文の候補数別頻度を示す。

表 2　実験対象動詞（64 語）

No.	かな表記	漢字表記	多義数
1	かえる	換える/帰る/替える/代える/変える/買える/返る	7
2	かえ	飼え/買え/換え/替え/代え/変え	6
3	たつ	経つ/建つ/絶つ/断つ/発つ/立つ	6
4	かける	掛ける/駆ける/欠ける/賭ける/書ける	5
5	とる	採る/撮る/執る/取る/捕る	5
6	さす	差す/刺す/指す/射す	4
7	しめる	絞める/占める/締める/閉める	4
8	あげる	挙げる/上げる/揚げる	3
9	うつす	移す/映す/写す	3
10	うつる	移る/映る/写る	3
11	おえ	終え/追え/負え	3
12	たえ	堪え/絶え/耐え	3
13	あやまる	誤る/謝る	2
14	かわる	代わる/変わる	2
15	せめる	攻める/責める	2
16	とける	解ける/溶ける	2
17	にる	似る/煮る	2
18	のぞむ	望む/臨む	2

表3　実験1試験文例

文番号	かな表記	漢字表記	試験文
15	うつす	移す	旗をラバウルへ（うつす）
150	うつす	写す	自分を 鏡に（うつす）
285	にる	煮る	肉をコンロで（にる）
325	にる	似る	犬は 飼い主に（にる）
1655	とる	採る	贈与の 形式を（とる）
1802	とる	撮る	写真を テープで（とる）
2322	かける	賭ける	上客が 金を カジノで（かける）
2395	かける	掛ける	問題の 混迷に 拍車を（かける）
4002	さす	射す	月影が 雲から（さす）
4019	さす	刺す	南京虫が 皮膚を（さす）

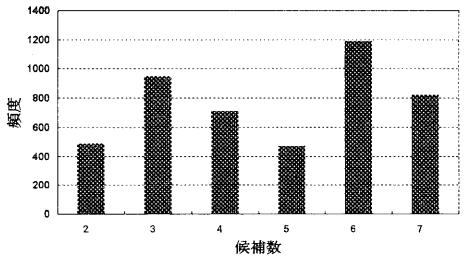

図2　候補数別の試験文頻度（動詞）

この試験文に対し、結合価パターンを適用し、変換を行う。

実際に動詞の同音異義語を選択する手順を以下の例文で示す。図 3 は、この例で使用する一般名詞意味属性体系の一部である。

入力文 N:連中が ロンドンを《たつ》

《たつ》の同音異義語:発つ、立つ

「連中」の意味属性:(人間)

「ロンドン」の意味属性:(都市)

結合価パターン X:(人)が (場所)を発つ

結合価パターン Y:(人)が (食料)を断つ

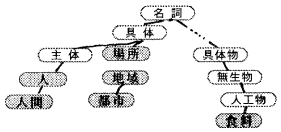

図 3　意味属性体系（一部）

(I)　入力文の未変換のかな動詞が同音異義語となるとき、その「かな文字列」を検索キーとして辞書から調べ、候補動詞とし、その動詞に対応する結合価パターンを調べる。

動詞《たつ》の結合価パターン→ X,Y

(II)　結合価パターンの格要素に対応する入力文の格要素の名詞に対して、辞書より意味属性を調べる。

助詞 1「が」→ X1:(人),Y1:(人)と N1:(人間)

助詞 2「を」→ X2:(場所),Y2:(食料)と N2:(都市)

(III)　結合価パターン側の意味属性が入力文側の意味属性の上位属性にあるか対応する格要素ごとに調べる。

結合価パターン X:X1 と N1→ OK, X2 と N2→ OK

結合価パターン Y:Y1 と N1→ OK, Y2 と N2→ NG

(IV)　それぞれの格要素の一致具合から、最も

一致する結合価パターンから、動詞を選ぶ。

結合価パターンX:2つの格要素が一致
結合価パターンY:1つの格要素が一致
↓
結合価パターンXを採用 →「発つ」を選択

以上のように、同じ読みの動詞を持つ各結合価パターンを入力文に適用し、最も一致した結合価パターンを調べることで、同音異義語の動詞を選択する。

実験2：名詞の同音異義語を選択する実験

名詞の実験では、始めに、計算機用日本語基本名詞辞書IPAL[4]（以下、IPAL名詞辞書）に収録されている基本名詞（1,081語）を、ALT単語辞書に収録されている全ての名詞（約40万語）と比較する。IPAL名詞辞書の中でALT単語辞書側の名詞と同音異義語の関係を取るのは500種519語存在する（表4）。このIPAL側の名詞に付随して収録されている用例文を、単文に直し整理したものを名詞用の試験文（504文）として用いる（表5）。図4に試験文の候補数別頻度を示す。

表4　実験対象名詞の一部

No.	かな表記	漢字表記	多義数
1	かん	冠/寒/勘/巻/完/刊/…	26
2	きこう	機構/奇行/寄港/寄航/寄稿/帰校/…	20
3	こうこう	後攻/航行/高校/口腔/坑口/孝行/…	19
4	きかん	器官/基幹/旗艦/帰還/既刊/期間/…	17
5	こうか	降下/高価/高架/硬化/効果/硬貨/…	17
6	かき	夏期/下記/夏期/夏季/火器/柿/…	16
36	せんしゅ	先取/先守/専守/船首/船主/選手/…	9
70	やく	ヤク/益/厄/役/約/訳	7
71	ようせい	幼生/妖星/妖精/陽性/要請/養成	7
118	でんき	伝奇/伝記/電器/電気/電機	5
281	みかん	未刊/未完/蜜柑	3
291	あと	後/跡	2
294	いいん	委員/医院	2
295	いきがい	域外/生き甲斐	2
344	きょうじゅ	享受/教授	2
392	しろ	城/白	2
448	にんき	人気/任期	2
484	みち	道/未知	2

表5　実験2試験文例

文番号	かな表記	漢字表記	試験文
8	きゅうしゅう	吸収	ものは(きゅうしゅう)が 早い
50	きょうか	強化	都内では 警備の(きょうか)が 続く
79	きょうりょく	強力	力を(きょうりょく)に する
188	きらい	嫌い	僕は チーズが(きらい)に 成る
250	いき	息	祖父は(いき)を 引き取る
252	いき	息	二人は(いき)が 合う
383	あんしん	安心	人は(あんしん)を 求める
405	とうばん	当番	(とうばん)を 済ませる
460	てんじょう	天井	彼は(てんじょう)を 見上げる
477	じょうしゃ	乗車	運転手は 酔客の(じょうしゃ)を 拒否する

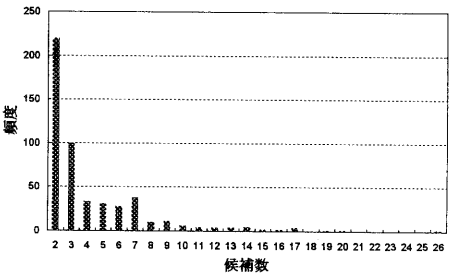

図4　候補数別の試験文頻度(名詞)

この試験文に対し、結合価パターンを適用し、変換を行う。

実際に名詞の同音異義語を選択する手順を以下の例文で示す。図5は、この例で使用する一般名詞意味属性体系の一部である。

名詞の同音異義語を変換する手順を、以下の例文で順に示す。

入力文M:信者に《きふ》を呼び掛ける

《きふ》の同音異義語:寄付、棋譜

「信者」の意味属性:(信徒)

「寄付」の意味属性:(商取引)

「棋譜」の意味属性:(出版物)

結合価パターンP:(主体)が(主体)に(行為)を 呼び掛ける

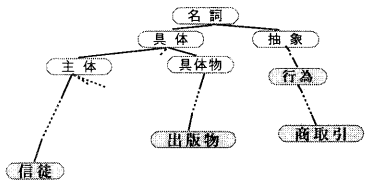

図5　意味属性体系（一部）

(I)　入力文の未変換の名詞が同音異義語とな

[4] 情報処理振興事業協会が作成した日本語辞書。

るとき、その名詞の読みと同じ語を調べ、候補名詞とする。

動詞《呼び掛ける》の結合価パターン→ P

(II) 入力文の動詞は変換済みと考え、その「表記」を検索キーとして辞書から対応する結合価パターンを調べる。

《きふ》の同音異義語:a)寄付、b)棋譜

(III) 実験 2 の手順 (II) と同様にそれぞれの意味属性を調べる。ただし、候補名詞を含む格要素は、候補名詞全てに対して意味属性を調べる。

助詞 1「が」→入力文側に無し
助詞 2「に」→P2:(主体)と M2:(信徒)
助詞 3「を」→P3:(行為)と
M3a:(商取引), M3b:(出版物)

(IV) 実験 1 の手順 (III) と同様に、上位下位関係を調べる。ただし、候補名詞を含む格要素は、候補名詞全てに対して調べる。

a)寄付:P2 と M2→OK, P3 と M3a→OK
b)棋譜:P2 と M2→OK, P3 と M3b→NG

(V) 手順 (I) から (IV) を残りの結合価パターンがあれば適用し、最も結合価パターンに一致した候補名詞を選ぶ。

a)寄付:2 つの格要素が一致
b)棋譜:1 つの格要素が一致
「寄付」を選択

以上のように、入力文の動詞に対応する結合価パターンを各候補名詞に対して適用し、最も一致した結合価パターンを調べることで、同音異義語の名詞を選択する。

なお、結合価パターンは、岩波書店「日本語語彙大系」の第 5 巻「構文体系」に掲載されている「構文意味辞書」から使用した。

(3) 変換結果の正誤判定

実験 1 では、抽出した試験文の表記そのものを正解表記とする。実験 2 では、IPAL 名詞辞書には、各用例ごとに使用され得る名詞が、「表記」の項目に収録されているので、この表記を正解表記とする。

3.2. 実験の結果

(1) 変換結果の分類

実験結果を次の基準で評価する。

実験結果を、以下の 6 つに分類した。

a) 「正解」: 選択候補数は 1 で、同音異義語の中から正解表記を選択できた場合
b) 「選択候補数減少」: 選択候補数は 2 以上で、正解表記を含む場合
c) 「誤答」: 選択候補の中に正解表記が含まれない場合
d) 「選択候補数不変」: 候補数と選択候補数が変わらない場合
e) 「選択候補なし」: 選択候補数が 0 の場合
f) 「パターンなし」: 試験文の動詞に対応する結合価パターンがなかった場合

さらに、上記「正解」と「選択候補数減少」の 2 つは、結合価パターンの「効果あり」とし、残りの 4 つは「効果なし」と分類する。

(2) 結果と考察

実験の結果は、実験 1, 実験 2 それぞれ、前述した分類による構成割合で示す(図 6,7)。「正解」は、実験 1 では 33%、実験 2 では 22%であり、「効果あり」は、49%と 37%であった。多義平均数は、動詞では 4.3 から 2.9 へ減少し、名詞では 4.2 から 3.3 へ減少した。また、同音異義語をランダムに選択した場合と比較

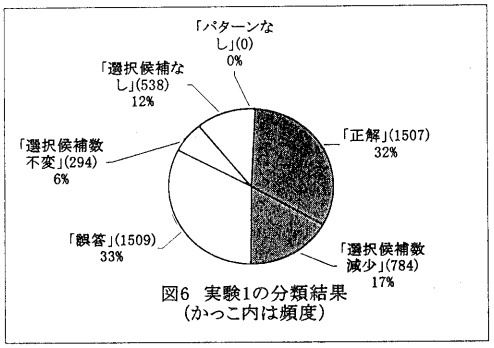

図6 実験1の分類結果
(かっこ内は頻度)

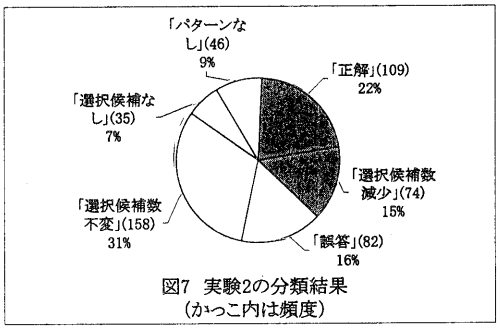

図7 実験2の分類結果
(かっこ内は頻度)

表6 選択できない例(上:実験1,下:実験2)

正解表記	出力	試験文
		(i) 格要素の省略、もしくは格要素が代名詞
挙げる	上げる	死を (あげる)
責める	責める/攻める	全員が (せめる)
指す	刺す	ものを 会社は (さす)
書ける	掛ける/駆ける	私が (かける)
		(ii) 上位属性を持つ結合価パターンの一致
代わる	変わる/代わる	風に (かわる)
代わる	変わる/代わる	政党に (かわる)

正解表記	出力	試験文
		(i) 格要素の省略、もしくは格要素が代名詞
最初	最初/細書	本の (さいしょ)は つまらない
木	機/気/木	(き)で 作る
		(ii) 上位属性を持つ結合価パターンの一致
麻	麻/朝	地方では 多くの (あさ)を 観察する
足	葦/足	彼女は (あし)が 長い
味	鯵	店の 料理は /味/(あじ)が 濃い
最後	最期/最後	映画は /最後/(さいご)が 面白い
		(iii) 必要な結合価パターンの不足
衣類	-	(いるい)は まとめて 荷作りする
辺り	-	あなたは (あたり)を わきまえる
上	-	壁の (うえ)に ポスターを 貼る
兄弟	-	彼の (きょうだい)が 事故死する

して、正解率が動詞では 25%から 48%に、名詞では 34%から 47%に上昇した。

同音異義語をうまく選択できない例を表 6 に示す。この原因として、以下のようなことが考えられる。

(i) 格要素の省略、もしくは格要素が代名詞

結合価パターンは、格要素の情報に依存するため、必要な格要素がないとパターンが一致しない。このような例は、多くの場合、人が見ても選択が困難であった。このような例では、前後の文から格要素を推定する必要があると考えられる。

(ii) 上位属性を持つ結合価パターンの一致

例えば「変わる」と「代わる」は、ともに最上位の意味属性、つまり名詞全般を許容する結合価パターンを持つ。この場合は、どちらの結合価パターンにも一致し、選択できない。しかしながら、このような試験文の原文の前後を参照したところ、人間ならば選択できる例もあった。この場合は、前後の文脈を考慮する必要がある。

(iii) 必要な結合価パターンの不足

分類 f)の「パターンなし」に該当するものである。これは、結合価パターンそのものが不足しているためで、特に、サ変動詞、複合動詞、名詞+「ダ」といった用言が不足していた。

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かな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を使用し、同音異義となる動詞を含む単文と、同音異義となる名詞を含む単文それぞれに対し実験を行った。この結果、同音異義語の中から正解表記を決定できた文は、動詞の場合で 33%、名詞の場合では 22%であった。候補数が減少した文は、それぞれ 49%と 37%であった。また、

絞り込めなかった候補をランダムに選択すると仮定し、全体をランダムに選択する場合と比較すると、正解率が動詞では 25%から48%に、名詞では 34%から 47%に向上した。

本論文の実験結果からは、格フレームを用いた従来の方法と比較して、ほぼ同程度の効果しか示せなかった。同音異義語の選択ができない理由は、結合価パターンの持つ意味的な制約だけでは、曖昧性が残り、同音異義語の候補を絞りきれないためと考えられる。

また、本論文で用いた結合価パターンは、日英機械翻訳における漢字かな混じり文の解析を目的に作成されているため、かな漢字変換の解析においては不十分である結合価パターンがあると考えられる。このため、同音異義語の選択という観点を考慮した結合価パターンの見直しが必要と思われる。

本論文では、結合価パターンのみで同音異義語の選択を行ったが、単語ごとに方法を変える、または他の方法と組み合わせて使うことで、変換精度が向上する可能性も考えられる。

参考文献

[1] 高橋,吉村,首藤(1995) : 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