結合価パターンを用いた仮名漢字変換候補の選択

吉田 真司　徳久 雅人　村上 仁一　池原 悟
鳥取大学 工学部 知能情報工学科
{yosita,tokuhisa,murakami,ikehara}@ike.tottori-u.ac.jp

1　はじめに

現在，仮名漢字変換において高い変換精度を誇る手法として，単語連鎖確率を用いた手法 [1],[2] がある．しかし，単語間の局所的な関係しか考慮しておらず，同音異義語の仮名漢字変換での誤変換が問題となる．

これに対し，文の意味解析を用いた仮名漢字変換の手法としては，連語共起情報を意味素で記述しその関係を調べる方法 [3]，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 [4] などが挙げられる．ただし，意味素の体系を適切に設定することや，一貫性のある格フレームを大規模に構築することが困難であった．しかし，近年，網羅性の高い意味的辞書として結合価パターン [5] が作成されたことにより，意味解析を用いる手法が現実味を帯びてきた．

そこで，本稿では，単語連鎖確率を用いた仮名漢字変換の候補文に対して，結合価パターンを用いた候補選択を行う手法を実現し，その有効性を調査する．

2　仮名漢字変換アルゴリズム

2.1　アルゴリズムの概要

単語連鎖確率を用いた仮名漢字変換は，確からしい順に変換候補を大量に出力することができる．変換候補に対し，結合価パターンを用いることで，意味的な適切性が判定できる．よって，両者を統合したアルゴリズムを作成する．アルゴリズムの概要を以下に示す．

(手順 1：候補作成)

(手順 2：文法的に不適切な文の削除)

(手順 3：意味的に不適切な文の削除)

(手順 4：慣用表現の優先)

(手順 5：補欠選択)

ここで，補欠選択とは，(手順 2) および (手順 3) で全ての候補が削除された場合に，出現確率の最も高い候補文を選択することである．以下に手順の詳細を示す．

2.2　単語連鎖確率を用いた変換候補作成

(手順 1) では，文献 [2] の単語連鎖確率を用いた仮名漢字変換を行う．

具体的にはまず，入力文を単語列に分割する．そして，単語列で隣接する単語間の連鎖確率を計算し，その組み合せから出現確率を求める．次に出現確率の高い文から順に，候補として出力する．

最後に，入力された平仮名文に対して，単語連鎖確率を用いた仮名漢字変換を行い，出現確率の高い上位 32 文の変換候補を作成する．

ここで，日本語語彙大系から約 35 万語の単語辞書を，EDR コーパス約 20 万文から接続辞書をそれぞれ作成した．

2.3　文法的に不適切な文の削除

(手順 2) では，ALT-JAWS を用いた形態素解析を行う．解析エラー時の出力を利用して品詞間の接続関係に誤りのある候補文を削除する．

2.4　意味的に不適切な文の削除

2.4.1　削除条件

(手順 3) において，候補を採用する条件は「候補文の全ての格要素と用言の間の意味的結束性が正しいこと」である．一部でも格要素の結束性が確認できない候補文は候補から削除する．

ところで，実際の文には結合価パターンの制約を受けない任意格が存在するため，結合価パターンだけでは，全ての格要素の結束性が確認できない．さらに，本稿の現状では，格要素を構成する名詞が名詞句である場合，名詞句の主名詞にしか一般名詞意味属性による制約が働かない．

そこで，本稿では，任意格を判定する為に，任意格のルールベースを用いる．名詞句の問題については，単語辞書，および，接続辞書を用いて，確認をとる．

任意格のルールベースは，例文より任意格要素を集めて作成する．任意格要素を判定するルールは，名詞部分を一般名詞意味属性で制約し，かつ，使用可能な格助詞を指定する．

2.4.2 意味的に不適切な文の判定

以下の段階を経て意味的に不適切な文を候補から削除する．

(手順 3-1．態による格助詞の変換)
結合価パターンは能動態で定義される．候補文が能動態でない場合，文献 [6] の格助詞変換規則に従い，パターンマッチングの為の格助詞を補足する．

(手順 3-2．結合価パターンのマッチング)
まず候補文の用言で結合価パターンを [5] より検索する．次に結合価パターンの規定する格助詞と一致する格要素を候補文から全て抽出し，格要素への意味属性制約を検査する．ここで，格助詞が一致するにも関わらず，意味属性制約が満たされないならば，その結合価パターンは適用しない．なお，結合価パターンの規定する格助詞が候補文で必ずしも使われる必要はない．マッチする結合価パターンが見つからないならば，候補文を削除する．

(手順 3-3．任意格の検査)
(手順 3-2) の段階で検査されなかった格要素は，任意格ルールベースで検査する．不適切な格要素が 1 つでもある場合，候補文を削除する．任意格ルールベースの例としては,「(388 場所) ＋『で格』」,「(2670 時間) ＋『に格』」などがある．

(手順 3-4．名詞句の検査)
(手順 3-2) および (手順 3-3) の段階で検査された格要素が名詞句ならば，名詞句を構成する単語間の接続関係を接続辞書により確認する．未知の接続を含む名詞句であるならば，候補文を削除する．

2.5 慣用表現の優先

結合価パターンには一般表現と慣用表現があり，一般表現は意味属性で，慣用表現は字面で，それぞれ格要素に制約を発生させる．

(例) 一般表現：$N1$(3 主体) が $N2$ (4 人) を絞る
慣用表現：$N1$(4 人) が 知恵 を絞る

日常的な文では，慣用表現の読みと同じになるような漢字の使い方はしないのが普通であるので，慣用表現のパターンが当てはまる候補文は，優先的に選択することが妥当である．

3 評価実験

3.1 評価実験の目的と方法

本研究では，以下の 3 つの実験を行う．

(実験 1) 本手法の動作の確認，および，任意格のデータベース作成を目的とする．EDR コーパスの単文集を対象とする．

(実験 2) 同音異義語に対する結合価パターンの効果の調査を目的とする．IPAL 動詞辞書，および，IPAL 名詞辞書に登録された，多義のある基本動詞，および，基本名詞から作成した単文集を対象とする．

(実験 3) 実際の文に対する結合価パターンの効果の調査を目的とする．毎日新聞 95 年度記事 [7] の単文集を対象とする．

ただし，本稿では未知語が含まれる文を対象外とする．よって，単語辞書に含まれている単語で構成された例文 100 文を用いる．

3.2 評価基準

評価は,「正解文との完全一致を正解とする基準」，および,「人手で正解文と見比べて判定する基準」の 2 つを設ける．人手の判定を設けた理由は，仮名漢字の表記の正しさが絶対的には定まらないこと，および，語義選択が元々不可能である場合があることに対処するためである．例えば，人手での判定によると,「復習に時間を掛ける」と「復讐に時間を掛ける」はいずれも正解となる．

3.3 実験結果

実験結果はアルゴリズムの各段階での効果を比べるため，手順ごとに N-best で正解率を集計する．実験 1 から実験 3 の結果を表 1 から表 4 に示す．各手順の上段は完全一致，下段は人手による判定の結果である．

実験 1 の「(手順 1) のみ」の結果より，(手順 1) の動作が正常に行われたことがわかる．

実験 2 の「同音異義語の動詞を含む文」での結果において,「(手順 1) のみ」と「(手順 4) まで」の結果の比較では，正解率に多少の向上が見られる.「全手順」を行った場合は，格段に正解率が上昇している．

実験 2 の「同音異義語の名詞を含む文」での結果では，手順を踏まえるに従って正解率が上昇している．この結果より，本手法は同音異義語の名詞に特に効果があることがわかる．

実験 3 では,「全手順」を行った場合，正解率が向上している．よって，本手法は実際の文に対しても有効であることがわかる．しかし,「(手順 4) まで」の結果を見ると,「(手順 1) のみ」の結果よりも正解率が低かった．次章では，この理由について考察する．

表 1：実験 1 の結果 (クローズド テスト)

手順＼候補数		1 位	～4 位	～16 位	～32 位
(手順 1) のみ	完全一致	83%	95%	98%	98%
	人手判定	96%	98%	100%	100%
(手順 2) まで	完全一致	78%	91%	93%	93%
	人手判定	95%	97%	99%	99%
(手順 4) まで	完全一致	66%	79%	80%	80%
	人手判定	83%	85%	85%	85%
全手順	完全一致	79%	93%	94%	94%
	人手判定	96%	99%	99%	99%

表 2：実験 2 の結果 (同音異義語の動詞を含む文)

手順＼候補数		1 位	～4 位	～16 位	～32 位
(手順 1) のみ	完全一致	34%	55%	72%	73%
	人手判定	55%	74%	88%	88%
(手順 2) まで	完全一致	35%	56%	70%	71%
	人手判定	54%	74%	84%	84%
(手順 4) まで	完全一致	35%	50%	54%	54%
	人手判定	58%	65%	67%	67%
全手順	完全一致	44%	62%	68%	68%
	人手判定	67%	78%	83%	83%

表 3：実験 2 の結果 (同音異義語の名詞を含む文)

手順＼候補数		1 位	～4 位	～16 位	～32 位
(手順 1) のみ	完全一致	36%	63%	77%	82%
	人手判定	58%	83%	94%	98%
(手順 2) まで	完全一致	37%	64%	78%	82%
	人手判定	61%	84%	95%	98%
(手順 4) まで	完全一致	37%	60%	66%	67%
	人手判定	67%	80%	82%	83%
全手順	完全一致	42%	69%	79%	81%
	人手判定	73%	89%	95%	97%

表 4：実験 3 の結果 (毎日新聞記事)

手順＼候補数		1 位	～4 位	～16 位	～32 位
(手順 1) のみ	完全一致	52%	76%	82%	87%
	人手判定	64%	79%	86%	91%
(手順 2) まで	完全一致	52%	77%	83%	86%
	人手判定	63%	79%	86%	89%
(手順 4) まで	完全一致	40%	57%	57%	57%
	人手判定	55%	60%	60%	60%
全手順	完全一致	58%	77%	78%	81%
	人手判定	73%	81%	83%	86%

4 考察

実験 3 の人手判定において,「(手順 4) まで」により 32 個の候補文を全て削除した件数の割合は 36%であった．したがって,「(手順 4) まで」は 64%の判定結果を出力したことになり，候補を正しく選択することの適合率 (⟨ 正解数 ⟩ / ⟨ 出力数 ⟩) は 86%である．これは「(手順 1) のみ」の適合率 64%よりも高い．

一方，再現率は,「～32 位」までをみても,「(手順 4) まで」は 60%であるが,「(手順 1) のみ」は 91%である．

これらの関係から，2 章で示した「単語連鎖確率による候補作成」と「統語的・意味的判断による候補選択」の統合アルゴリズムは全体において両者の利点が活かされていると，定量的に結論づけられる．

そこで,「(手順 4)」の適合率と再現率の向上に向け，誤り事例の分析を行う．

4.1 誤り文が削除できないことについて

「(手順 4) まで」における「1 位」と「～32 位」の正解率の差より，誤り文が削除できなかった量がわかる．原因，および，該当候補文の数を表 5 に示す．

表 5：誤り文を削除できなかった原因

原因	文数
(1-1) 結合価パターンが原因	2
(1-2) 結合価パターン以外が原因	3

以下に具体例を載せる．

(原因 1-1) の例

入力文： しきちないにちいさなばばがある。
正解文： 敷地内に小さな馬場がある。
誤り候補文： 敷地内に小さな 祖母がある。

次のパターンで判断したため誤った．

パターン： $N1$(*) が $N2$(388 場所 533 具体物 1000 抽象) にある

この場合,「祖母 (人間)」に対して「ある」は誤りである．

(原因 1-2) の例

入力文： てんしんのことばがよみがえる。
正解文： 転進の言葉がよみがえる。
誤り候補文： 点心の言葉 がよみがえる。
パターン： $N1$(*) が蘇る

誤り候補文において「点心」と「言葉」は，人手での評価では繋がらないと判断し，誤りとした．

以上より，(原因 1-1) の場合，結合価パターンの制約条件の変更が必要となる．しかし,「『彼が～の立場にある』という場合では正しい」というように，一般に制約条件の変更は容易ではない．(原因 1-2) は，人手での判定では誤りとしたが,「の」型名詞句は，状況や考え方によっては繋がることがある．したがって，これら 2 つの原因の解決は困難である．

4.2 正解文が削除されることについて

(手順 2) における「～32 位」と (手順 4) における「～32 位」の正解率の差より，正解するはずの候補文が，意味的削除で削除されてしまった量がわかる．その原因を，実験 3 の結果より調べ，表 6 にまとめる．

表 6：正解文が削除された原因

原因	件数
(2-1) 結合価パターンの名詞意味属性の制約が不足	10
(2-2) 任意格データベースに登録された名詞意味属性の制約が不足	7
(2-3) 結合価パターンの不足	5
(2-4) 接続辞書のデータ不足	4
(2-5) 格変化規則が不適切	3

以下に具体例を載せる．

(原因 2-1) の例

正解文： 国籍 を拒否された。
パターン： $N1$(3 主体) が $N2$(3 主体 1236 人間活動) を拒否する

「国籍」の意味属性は「(1203 籍)」であり，(3 主体 1236 人間活動) の意味属性とは親子関係にないため，パターンが適合せず，正解文が削除された．

(原因 2-2) の例

正解文： これまで、同公社内部に 保管されていた。
パターン： $N1$(3 主体) が $N2$(533 具体物 1001 抽象物) を保管する

正解文には，任意格の「に格」が含まれている．しかし，任意格のデータベース中に「内部 (2623 内部) に」に対応する意味属性が無いので，正解文が削除された．

(原因 2-3) の例

正解文： 機長と副操縦士のやりとりが 生々しい。
パターン： なし

「生々しい」は，単語辞書には登録されているが，結合価パターン辞書には「生々しい」が存在しない．よって，正解文が削除された．

(原因 2-4) の例

正解文： 派遣予定 もない。

「派遣予定」の名詞の接続において「派遣」と「予定」の接続が接続辞書に載っていないため，正解文が削除された．

(原因 2-5) の例

正解文： 同氏自身の具体的な 見解 は示さなかった。
パターン： $N1$(3 主体) が $N2$(*) を $N3$(1109 文書 1062 言語) に示す

「見解は」の対応する部分は「N2(*) を」なので，「は格」を「を格」に変化させなくてはならない．しかし，格変化規則では,「は格」を「が格」に変換することが優先されるため，結合価パターンが誤って適合したことが原因で，正解文が削除された.「が格」を優先した理由は「は格」を「が格」，および,「を格」の両方と対応させた場合，誤った判断をすることが動作確認の際に多く見られたためである．例えば,「私は作る」の誤り文で「渡しは作る」という文がある．この場合,「を格」を用いると「渡し（橋）を作る」と解釈されてしまう．そのため，本稿では使用を控えている．

以上より，(原因 2-1)〜(原因 2-4) ではデータベースの拡充，(原因 2-5) では格変化規則の改良がそれぞれ必要となる．これらについては改良の余地があるため，今後の精度向上が期待できる．

5 おわりに

本稿では，単語連鎖確率を用いた仮名漢字変換の候補文に対して，結合価パターンを用いた候補選択を行うという統合的な仮名漢字変換アルゴリズムを提案した．IPAL の基本動詞/基本名詞に関する同音異義語を含む仮名文，および，毎日新聞 95 年度記事の仮名文それぞれ 100 文を対象とした仮名漢字変換の評価実験では，1 位正解率については，それぞれ 12%，15%，および 9%の向上が見られた．よって，本アルゴリズムの有効性が確認できた．今後の課題は，データベースの拡充，および，格変化規則の改良である．

謝辞

ALT-JAWS は NTT との共同研究の下で使用させて頂きました．

参考文献

[1] 村上: 漢字かなの TRIGRAM をもちいたかな漢字変換方法, 情報処理学会第 43 回全国大会, 7H-3, Vol.3, pp.287-288(1991).

[2] 森, ほか 3 名: 確率的モデルによる仮名漢字変換, 情報処理学会論文誌, Vol.40, No.7, pp.2946-2953(1999).

[3] 本間, 山階, 小橋: 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(1986).

[4] 大島, ほか 3 名: 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(1986).

[5] 池原, ほか 7 名: 日本語語彙大系, 岩波書店 (1997).

[6] 吉田, 池原, 村上: 入力文に対する結合価パターン対の選択方法について, 言語処理学会第 8 回年次大会発表論文集, pp.299-302(2001).

[7] 日外アソシエーツ株式会社: 毎日新聞 95 年度版 CD-ROM(1995).