格フレームを用いたかな表記語の曖昧性解消

岡部 浩司　　河原 大輔　　黒橋 禎夫
東京大学大学院情報理工学系研究科
{okabe, kawahara, kuro}@kc.t.u-tokyo.ac.jp

1 はじめに

日本語では，漢字表記をもつ単語にもしばしばかな表記が用いられる．それにより，異なる漢字表記をもつ同音異義語が同一のかな表記で表され，曖昧性が生じるという問題がある．例えば,「かぜを引く」という文で,「かぜ」というかな表記語が，漢字表記では「風」と「風邪」のどちらであるかという曖昧性が生じる．

SENSEVAL-2[3] など，語義曖昧性解消については多くの研究がなされてきたが，書き言葉ではかな表記はあまり使われないことから，かな表記語の曖昧性解消については多くの研究はなされていなかった．かな漢字変換では，かな表記語の曖昧性解消を行っており，頻出する語については対応できているものの，Web テキストなどではくだけた文章が多く，様々な語がかな表記されるため，対応できないものも多い．

形態素解析器 JUMAN では，代表表記という概念を設け，かな表記や送りがなによる同じ語の表記バリエーションを一つの代表表記で扱うことが可能である．さらに，辞書の語彙を現代日本語で標準的に用いられる約 3 万語に限定したため，かな表記語による曖昧性があった場合も，その中から複数の候補が挙げられるようになっており，かな表記語の曖昧性解消に着手することができるようになった．

我々がこのようなかな表記語における曖昧性を解消して理解することができるのは，互いに係り受けの関係にある名詞と用言が，実際にその漢字表記で使われるかどうかを知っているからである．前述の例で言えば,「引く」という動詞を「風邪を引く」という文で使うことはあるが,「風を引く」という文で使うことはないという知識をもっているため，曖昧性を解消することができると考えられる．

本稿では，このような知識として，コーパスから自動構築された格フレーム辞書を用い，かな表記語の曖昧性を解消することを試みる．用言においては，格構造の類似度から格フレームを決定することにより曖昧性を解消し，名詞においては，係り先の用言の格フレームを参照することにより曖昧性解消を行う．本手法によってかな表記語の曖昧性解消実験を行い，その結果を考察する．

```
かぜ かぜ かぜ 名詞 6 普通名詞 1 * 0 * 0 "漢字読み:訓 代表表記:風/かぜ"
@ かぜ かぜ かぜ 名詞 6 普通名詞 1 * 0 * 0 "代表表記:風邪/かぜ"
で で で 助詞 9 格助詞 1 * 0 * 0 NIL
おくれた おくれた おくれる 動詞 2 * 0 母音動詞 1 タ形 8 "可能動詞:送る 代表表記:送れる/おくれる"
@ おくれた おくれた おくれる 動詞 2 * 0 母音動詞 1 タ形 8 "付属動詞候補（基本） 代表表記:遅れる/おくれる"
EOS
```

図 1: JUMAN の解析例

2 代表表記

形態素解析器 JUMAN では，同じ語の表記バリエーションを扱うために，代表表記を設けており，これを意味情報として出力できる．これにより表記揺れの問題を，形態素解析を行うだけである程度取り除くことが可能である．

かな表記等による曖昧性がある場合は，日常の使用の範囲で複数の可能性（代表表記）を挙げるようにしている．例えば,「ふきん」には「付近」,「附近」,「布巾」,「賦金」,「斧斤」という漢字表記が存在するが,「付近」と「附近」は一つの代表表記でマージされ,「賦金」と「斧斤」は削除されており,「付近/ ふきん」,「布巾/ ふきん」の二つの代表表記が候補として挙げられる．その複数の代表表記から一つを決定することで曖昧性を解消することが可能となる．したがって，JUMAN の代表表記の整備によって，かな表記語の曖昧性解消を行うことが可能となったといえる．

図 1 は「かぜでおくれた」という文を JUMAN により形態素解析した結果である．曖昧性のある語は@の後に複数個の候補が出力される．これらの候補の中

から一つの代表表記を選ぶことを本研究でのタスクとする．

3　格フレーム辞書

格フレーム辞書は [1]，[2] で自動的に構築されたものを用いた．本節では格フレーム辞書の構築方法について概略を述べる．

格フレーム辞書構築の手順を以下に示す．

1. テキストを構文解析する．
2. 構文解析結果から信頼度の高い述語項構造を抽出する．
3. 抽出した述語項構造を用言とその直前の格要素ごとにまとめ，(最初の) 格フレームをつくる．以後，用言の直前の格要素を「直前格要素」，その格を「直前格」と呼ぶ．
4. 3でつくった格フレームをシソーラスに基づいてクラスタリングし，類似しているものをマージする．
5. 格フレームごとに必須格を選択する．直前格の用例数に対して，閾値以上の用例をもつ格を必須格とする．ただし，ガ格は常に必須格とする．

4のクラスタリングでは，シソーラスの根からの階層の深さと，一致する階層の深さを用いて定義された類似度を利用して行っている．

本研究では，日本語 Web ページ約 1 億ページから収集された，約 5 億日本語文のテキストから構築された格フレーム辞書を用いた．Web テキストは新聞テキストなどと異なり，ドメインの偏りが少なく，よりカバレージの高い，常識的な知識を多く含む格フレームを構築することができる．

現時点の格フレーム辞書は，用言と格要素の両方ともコーパスに存在する表記をそのまま用いて構築している．現在は代表表記を用いて構築する作業を行っており，それによってより実用的な格フレーム辞書となるだろう．

4　曖昧性解消の手順

かな表記語の曖昧性には, 用言曖昧性と名詞曖昧性の二種類が存在する．これらはかな表記されている単語がそれぞれ用言と名詞である．本節ではこの二種類に対して，それぞれの曖昧性を解消する手順を述べる．曖昧性解消は JUMAN によって曖昧だと判断され複数の候補が出力された語に対して行われ，用言とその格構造を入力とする．

表 1: 「むく」の格フレーム

用言	格	用例
向く：1	ガ	気，方，人，…
	ヲ	方，方向，前，…
	ニ	人，方向，方，…
向く：2	ガ	顔，頭，…
	ヲ	下，上，…
	ニ	＜補文＞，声，言葉，…
…		
剥く：1	ガ	女性，私，…
	ヲ	皮，リンゴ，栗，…
	ニ	＜補文＞，側，状，…
	ノ	ジャガイモ，リンゴ，…
剥く：2	ガ	獣，＜数量＞，波，…
	ヲ	牙
	ニ	人間，私，我々，…
…		

4.1　用言曖昧性解消

用言曖昧性解消は，漢字表記された用言の格フレームを選択することで行われる．格フレームの選択方法を例とともに示す．

(1) 皮をむく

という文を解析すると,「むく」には「向く」と「剥く」の二種類の漢字表記があり，曖昧性がある．この「むく」の格フレームを選択することで，曖昧性を解消する．

「むく」には表 1 のような格フレームが存在する．「皮をむく」は直前格がヲ格であるため，格フレームのヲ格の用例群と直前格要素である「皮」との類似度を計算する.「剥く：1」のヲ格の「皮」とマッチし，類似度が最大となるため，格フレームは「剥く：1」に決定される．こうして,「むく」の表記を「剥く」に決定する．

格フレームの選択法の詳細を示す．まず，以下の条件を満たした場合に格フレームの選択を行う．

1. 入力側の対象用言が直前格要素 C をもつ．
2. 直前格要素 C と直前格 cm が以下のいずれかの条件を満たす．
   - cm がヲ格，二格のいずれかである．
   - cm がヲ格，二格以外で，C が意味属性＜主体＞をもたない．
3. cm をもつ格フレームが存在し，cm の用例群と C の類似度が閾値以上である．

表 2: 「食べる」の格フレームのヲ格

用言	格	用例
食べる	ヲ	ご飯，ラーメン，…，そば，…，蕎麦，…

条件 3 を満たす格フレームのなかで，もっとも類似度が高い格フレームを選択する．ここで用いる類似度とは直前格要素と格フレームの類似度のうちもっとも高いものとする．用例間の類似度はシソーラスを用いて計算する．また類似度が同点の場合は，直前格の頻度が多い格フレームに決定する．

また，用言の曖昧性は「ぬって (塗って，縫って)」のように，活用することで現われるものもあることを付け加えておく．

4.2 名詞曖昧性解消

名詞曖昧性解消は，名詞が係る用言の格フレームを参照し，JUMAN によって出力された複数個の代表表記の中から一つを選択することで行われる．用言の格フレームにおいて，名詞の格と同じ格の格要素を参照する．

具体例として，

(2) そばを食べる

という文を考えよう.「そば」には「傍」と「蕎麦」の二種類の漢字表記があり，曖昧性が存在する．

まず「食べる」の格フレームが，直前格であるヲ格の「そば」によって決定される．決定された格フレームのヲ格を表 2 に示す．まず，名詞の漢字表記候補が格フレーム中に存在すれば，その漢字表記に決定する．つまり,「食べる」の格フレームの格要素のなかに,「傍」または「蕎麦」が存在するかどうかが調べられる．この例では「蕎麦」が格フレーム中にあるため，表記が「蕎麦」に決定され，曖昧性が解消される．もし，格フレームに「蕎麦」が存在しなかった場合，つまり全ての漢字表記候補が格フレーム中に存在しない場合は，次のように類似度を用いて決定する.「傍」,「蕎麦」の両方の漢字表記候補に対して,「食べる」のヲ格の格要素全て (ただし,「そば」は除く) との類似度を計算し，類似度の最大値が大きかった方に決定する．類似度はシソーラスを用いて計算する.「ご飯」,「ラーメン」などと類似度が高いのは「蕎麦」の方であるから，表記が「蕎麦」に決定される．

表 3: 曖昧性解消実験の結果

Web テキスト 675 文		
名詞曖昧性解消	4/4	100.0%
用言曖昧性解消	30/40	75.0%
料理テキスト約 7,800 文		
名詞曖昧性解消	33/43	76.7%
用言曖昧性解消	252/330	76.4%

5 曖昧性解消実験

前節で述べた曖昧性解消手法を構文解析器 KNP に実装し，実際のテキストを用いた曖昧性解消実験を行った．使用したテキストは Web テキスト 675 文と，料理テキスト約 7,800 文である．これらの文を形態素解析器 JUMAN によって形態素解析した後，KNP で構文解析，曖昧性解消を行った．JUMAN によって正しく形態素解析されたものに対して，曖昧性が正しく解消できているかどうかを人手で判断した．曖昧性解消の精度を表 3 に示す．また，曖昧性解消の成功例と失敗例を表 4 に示す．上段が JUMAN の形態素解析によって得られたかな表記語の漢字表記候補であり，曖昧性解消によって決定された漢字表記が【 】で囲まれている．下段が解析対象の文であり，曖昧性をもつかな表記語が [ ] で囲まれている．

名詞，用言曖昧性解消の精度はどちらも約 75%であった．以下に主な誤り原因を示す．

名詞曖昧性解消

名詞曖昧性解消の誤りの一つに,「めんもユニーク」の例のように，格フレームからのみでは曖昧性の解消は困難だが，その直前の「具材も変わってるし」という文や「めん」が単独で使われていて「〜面」といった表現ではないといった情報，すなわち，文脈や，複合語，名詞格フレームといった情報を用いることで曖昧性の解消が可能だと考えられる例があった．曖昧性解消を行う際に，これらの情報も用いるようにすることで，より精度のよい曖昧性解消が実現できると考えられる．

用言曖昧性解消

用言曖昧性解消での主な誤り原因の一つは，格解析の誤りによるものであった．例を挙げると,「今いる場所」や「女性に人気のあった透明感」などである.「今いる場所」の場合,「いる」の正しい漢字表記は「居る」である．この文では「いる」が「場所」に連体修飾し

表 4: 曖昧性解消の成功例，失敗例

＜成功例＞

・名詞曖昧性解消

【壺】| 坪

子供達を叱ったはずみにミルクの［つぼ］を割ってしまうのですが、…

【布巾】| 付近

豆腐は［ふきん］に包んで水気をしぼりとります。

・用言曖昧性解消

【付ける】| 着ける | 漬ける | 点ける | 就ける

論文の題名の後にキーワードを５つ［つける］こと。

【塗る】| 縫う

マッサージのあと、塩を［ぬった］ままラップを当て、ヘアバンドなどで押さえて固定し、…。

＜失敗例＞

・名詞曖昧性解消

【面】| 綿 | 麺 | 免

武内「具材も変わってるし、［めん］もユニーク。それじゃあ行ってみましょうか」。

【ガン】| 癌 | 雁

ひいては大腸［ガン］の予防や糖尿病の治療にも役立つといわれるなど、…

・用言曖昧性解消

【合う】| 有る | 会う

『クレイン』は、合評会でも女性に人気の［あった］透明感のある味わいが特徴。

【駆ける】| 掛ける | 欠ける | 懸ける | 書ける | 賭ける | 架ける

南九州産黒豚肉を時間を［かけて］柔らかく煮こみました。

ており,「いる」と「場所」の関係は,「(〜が) 場所にいる」というニ格である．しかし，格解析の結果，これを「場所がいる」とガ格で扱っているため，曖昧性解消の結果,「居る」ではなく「要る」だと判断してしまう．同様に「女性に人気のあった透明感」では「の」は主格を表す「の」であり，ガ格のように扱うことができる．しかし，格解析ではこれをガ格と判断できていなかった．これらの高度な格の扱いに対しても検討する必要があろだろう．

また，前述のように，現在格フレーム辞書を代表表記を用いて構築中であり，より良質な格フレームが実現できるため，さらに精度良く曖昧性解消ができると考えている．

6 おわりに

本稿では，かな表記による曖昧性を，日本語 Web ページから自動的に構築された格フレーム辞書を用いて解決する手法を提案した．また，本手法を用いた曖昧性解消実験において，実際のテキストで 75%程度の精度で曖昧性を解消することができた．

今後は，文脈，複合語，名詞格フレームといった情報の追加や，格フレーム辞書の代表表記を用いた構築を行い，さらに精度のよい解析を目指す予定である．

参考文献

[1] 河原大輔, 黒橋禎夫, “用言と直前の格要素の組を単位とする格フレームの自動構築”, 自然言語処理, Vol.9, No.1, pp.3-19 (2002)

[2] 河原大輔, 黒橋禎夫, “高性能計算環を用いた Web からの大規模格フレーム構築”, 情報処理学会 自然言語処理研究会 171-12, (2006)

[3] 黒橋禎夫, 白井清昭, “SENSEVAL-2 日本語タスク”, 電子情報通信学会, 言語理解とコミュニケーション研究会 pp.1-8 (2001)