代表表記による自然言語リソースの整備

岡部 浩司
東京大学大学院
情報理工学系研究科
okabe@kc.t.u-tokyo.ac.jp

河原 大輔
独立行政法人
情報通信研究機構
dk@nict.go.jp

黒橋 禎夫
京都大学大学院
情報学研究科
kuro@i.kyoto-u.ac.jp

1 はじめに

日本語の言語処理の抱える問題の一つとして，同じ語が異なった表記で用いられる，表記揺れの問題が挙げられる．日本語は同じ単語が漢字表記とかな表記の両方で書かれるなど，同じ語に対して複数の表記の仕方がある言語である．さらに，かな表記では,「風」も「風邪」も同じ「かぜ」になることから曖昧性が生じるという問題もある．

従来の自然言語処理において，単語は表記そのもの，またはその原形によって区別されていた．したがって，表記揺れに関しては全く別の単語として扱われ，かな表記語に関しては，異なる語であるのに，同じかな表記語としてまとめられて扱われていた．

形態素解析器 JUMAN5.0 では代表表記を実装し，同じ語の表記揺れには，同じ代表表記を与えた [1]．これによりその後の解析では，代表表記で単語を扱うことによって，表記揺れを気にすることなく解析を行うことができるようになった．しかし，構文解析器 KNP では，構文解析，格解析の中でシソーラスや格フレーム辞書といった自然言語リソースを扱っているが，これらは表記をそのまま用いており，曖昧性のある語や表記揺れを含んだままの状態であるという問題があった．

本研究では，代表的な自然言語リソースであるシソーラス，格フレーム辞書の代表表記化を行い，これらが持っていた表記揺れ，かな表記による曖昧性を解消することを提案する．これにより，シソーラス，格フレーム辞書中の表記揺れや，かな表記による曖昧性が原因で起こる解析誤りを減少させ，解析の精度を上昇させることが目的である．

2 代表表記

代表表記とは，各語に対して与えられた ID であり，同一の語の表記揺れには同一の代表表記が与えられているため，互いに表記揺れであることが分かる．代表表記は，代表的な表記とその読みのペアで,「蛍/ほたる」のように表される．以下に代表表記によってまとめられる表記揺れの例を示す．

11030040201	仮想現実	かそうげんじつ	仮想/かそう+現実/げんじつ
11030040301	試し	ためし	試し/ためし v
11030050101	不思議	ふしぎ	不思議だ/ふしぎだ
11030050102	七不思議	ななふしぎ	七不思議/ななふしぎ
11030050103	神秘	しんぴ	神秘だ/しんぴだ
11030050104	ミステリー	みすてりい	ミステリー/みすてりー
11030050105	オカルト	おかると	オカルト/おかると
	(従来)		(代表表記化)

図 1: 分類語彙表

漢字とかな 蛍 ほたる ホタル → 蛍/ほたる
送りがな 表す 表わす → 表す/あらわす
漢字表記 奇跡 奇蹟 → 奇跡/きせき
カタカナ語 デジタル ディジタル → デジタル/でじたる

形態素解析器 JUMAN5.0 では代表表記を意味情報として出力でき，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，かな表記語の「ふきん」では,「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．

3 シソーラスの代表表記化

本研究で用いたシソーラスは分類語彙表であり，その一部が図 1 の左部分である．分類語彙表の各エントリーは，意味コード，見出し，読みから成る．

シソーラスによって単語の持つ意味コード（分類語彙表では 11 桁のコード）を得ることができるが，表記揺れによってシソーラスの記載と異なる表記をしている語は，シソーラスから意味コードを得ることができないという問題がある．例えば，分類語彙表には「ミステリー」という語のエントリーがあるが,「ミステリ」という語のエントリーはなく,「ミステリ」の意味コードを得ることができない．このような表記揺れに関する問題を解消するために，シソーラスの代表表記化を行った．シソーラスのエントリーが代表表記で記載されていれば,「ミステリー」と「ミステリ」という語はどちらも代表表記が「ミステリー/みすてりー」

であるから，どちらの語を入力としても同じ意味コードを得ることができる．

また，分類語彙表には，かな表記語による曖昧性を持ったエントリーが存在する．例えば，見出しが「あく」というエントリーでは,「灰汁」と「悪」のどちらの意味で記載されているかという曖昧性が生じている．このようなエントリーの正しい表記を整理する意味でも，代表表記化が必要である．

3.1 ルールによる自動的な代表表記変換

JUMAN の辞書と KNP による解析を用いて，自動的にシソーラスの各エントリーの代表表記化を行った．以下にそのルールを示す．

1. 分類語彙表の各エントリーに対して，JUMAN の辞書に表記と読みが一致する語があるかどうかを調べる．
   - JUMAN の辞書と一致する語はその語の代表表記を出力する．
     例）接続 せつぞく → 接続/せつぞく
     奇蹟 きせき → 奇跡/きせき
   - 複数と一致すれば各代表表記を?で連結し，すべて列挙する．
     例）そば そば → 傍/そば?蕎麦/そば
   - ナ形容詞，長音といった JUMAN の辞書と分類語彙表の表記の違いを考慮する．
     例）スタンダード すたんだあど
     → スタンダードだ/すたんだーどだ
2. 1 で JUMAN の辞書に一致する語が存在しない時，複合語や品詞変更の可能性を考え、見出し語を構文解析器 KNP で解析する.（品詞変更した語の代表表記を得るために KNP を用いる.）
   - KNP から出力された各語の読みをつなげ，分類語彙表の読みと比較する．一致した時，複合語とみなし，それぞれの代表表記を”+” で繋げたものを出力する.（1 と同様の処理も行う）
     例）粒あん つぶあん
     → 粒/つぶ+暗/あん?案/あん?餡/あん
   - 品詞変更や複合語の濁音化を考慮する．品詞変更した語は代表表記末尾に v が付けられる．
     例）たまり醤油 たまりじょうゆ
     → 溜まり/たまり v+醤油/しょうゆ
   - 代表表記を持たない語（地名，人名，組織名，固有名詞，数詞，連語等）は,「見出し/読み」の形で KNP が出力する疑似代表表記を出力する．
     例）アメリカ あめりか → アメリカ/あめりか
3. 1, 2 において代表表記が一つも挙げられないエントリーは，代表表記化した分類語彙表のエントリーから除く．
   例）殷鑑 いんかん → ×

3.2 人手による正しい代表表記の判断

ルールによって自動的に代表表記化した後，見出しにひらがなを含むエントリーに関しては，代表表記に?を含み曖昧性を持つことや，実際には JUMAN の辞書にない語なのに別の語の代表表記が列挙されることがある．例を挙げると，

(i) ねぎ ねぎ → 葱/ねぎ
(ii) 粒あん つぶあん
→ 粒/つぶ+暗/あん?案/あん?餡/あん

(i) のエントリーは本来「禰宜」の意味だが，この語は JUMAN の辞書に採用しておらず,「葱/ねぎ」のみが挙げられてしまっている．(ii) は「餡」の意味だが「暗/あん」,「案/あん」,「餡/あん」の三種類が列挙されてしまっている．

そこで，これらに対して人手で以下の作業を行った．

- 列挙されている中に正しい代表表記があれば，その代表表記に決定する．
- 列挙されている中に正しい代表表記がなければ，新たに JUMAN の辞書にその語を追加するか，JUMAN の辞書には含めず，分類語彙表にも用いないことにするかを判断する．

この作業により，かな表記の見出しを持つエントリーの曖昧性を解消することができ，シソーラスを用いた解析時に誤った意味コードを出力してしまうことを防ぐことができる．

以上によって代表表記化された分類語彙表の一部を図 1 の右部分に示す．

4 格フレーム辞書の代表表記化

格フレーム辞書は用言の用法を記述し，構文の曖昧性を解消するなどに役立つ．既存の格フレーム辞書は各語の原形を用いて構築されており，表記揺れや曖昧性を持った語が格フレーム辞書中に存在していた．

格フレーム辞書を代表表記化することで,「表す」,「表わす」といった用言の表記揺れを「表す/あらわす」にまとめ，格要素の「研鑽」,「研さん」も「研鑽/けんさん」にまとめることができる．

また,「あめ」という格要素は「雨」と「飴」の曖昧性を持つが，このような曖昧性を解消しつつ格フレームを構築する．さらに，かな表記の曖昧性により用例が不十分な用言があり，その補完も行う．

4.1 格フレーム辞書構築

格フレーム辞書の構築手法には河原ら [2] の手法を用いた．この中で用言，格要素を代表表記で扱い，類似度計算に代表表記化した分類語彙表を用いた．複合語は代表表記を”+” で連結し，曖昧性があるものは”?”

で代表表記列を連結する．各述語項構造を用例と呼び，用例を用言，直前格要素ごとにまとめたものを用例パターンと呼ぶ．

4.2 曖昧性を持つ格要素の処理

かな表記等によって格要素が曖昧性を持っている語は,「飴/あめ?雨/あめ」というように候補代表表記を列挙された形となっている．格フレーム構築の中で行われる，直前格要素の意味コードの固定，その他の格要素の意味コードの削除を用いて，これらの曖昧性の解消を試みる．

意味コードの固定

格フレーム構築の中で用言，直前格要素ごとに用例をまとめた用例パターンのクラスタリングを行うが，この際に直前格要素間の類似度を計算する．この時，多義である語は全ての意味コードで類似度を計算し，類似度が最大となるような意味コードを選択する．これを意味コードの固定と呼ぶ．曖昧性をもつかな表記では，この固定された意味コードに対応するものに代表表記を決定する．

例えば,「ガン/がん?癌/がん?雁/がん[1]」と「 鳥/とり」をそれぞれ直前格要素に持つ用例パターンをクラスタリングすると,「雁/がん」の意味コードを用いた際に類似度が最大となるため,「雁/がん」に代表表記を決定する．

意味コードの削除

直前格要素以外の格要素では，クラスタリングの後に他の格要素との類似度を計算する．この時，他の格要素との類似度が低い意味コードは用いないよう制限し，これを意味コードの削除と呼ぶ．曖昧性をもつかな表記の候補代表表記の意味コードが全て削除された場合，その代表表記を候補から除外する．

例を挙げると，ある格フレームの格要素が「オーブン/おーぶん」と「釜/かま? 鎌/かま?窯/かま」であったら,「釜/かま」や「鎌/かま」の持つ意味コードとは他の格要素との類似度が低いため，削除される．すると「釜/かま?鎌/かま?窯/かま」の持つ意味コードは「窯/かま」のものだけとなるので，代表表記を「窯/かま」に決定する．

4.3 曖昧性を持つ用言の格フレームの処理

代表表記化された格フレーム辞書では，従来では「躓く」,「つまずく」というように別々の用言として扱われていた用言の表記揺れを,「躓く/つまずく」という代表表記によってまとめている．これによって，例え「躓く」という漢字表記があまり用いられず用例が十分に集まらなくても,「つまずく」というかな表記の用例によって「躓く/つまずく」の用例を集めることが可能である．

しかし,「掻く/かく」の場合は漢字表記の「掻く」から十分に用例が集まらないからといって，上述のように「かく」の用例から単純に「掻く/かく」の用例を集めることはできない．これは「かく」というかな表記が,「欠く/かく」,「書く/ かく」,「掻く/かく」のいずれかであるという曖昧性を持つためである．このような曖昧性を持つ用言の格フレームを適切に分類し，各用言の用例を充足させる処理について述べる．

用言の分類

曖昧性を持つ用言（例．欠く/かく?書く/かく?掻く/かく）の用例パターンを曖昧用例パターン，曖昧性を持たない用言（例．欠く/かく）のものを一意用例パターンと呼ぶことにする．用例パターンが，一意用例パターンで集まるか，曖昧用例パターンで集まるかによって，用言を以下のように分類する．

I 全ての表記が曖昧性を持ち，曖昧用例パターンのみしか集まらない用言

I-i 漢字表記，かな表記のどちらも曖昧な用言

例) 弾く/ひく

（弾く→弾く/ひく?弾く/はじく，
ひく→引く/ひく?挽く/ひく?弾く/ひく）

I-ii 漢字表記がなく，かな表記が曖昧な用言

例) くれる/くれる

（くれる→くれる/くれる?暮れる/くれる）

II 一意用例パターンと曖昧用例パターンの両方が集まる用言

II-i かな表記が曖昧であり，かつ，かな表記される頻度が非常に高いため，一意用例パターンが十分に集まらない用言

例) 掻く/かく

（かく→欠く/かく?書く/かく?掻く/かく）

II-ii かな表記が曖昧だが，漢字表記される頻度が高く，一意用例パターンが集まる用言

例) 会う/あう

（あう→会う/あう?合う/あう）

III どの表記も曖昧性を持たず，一意用例パターンのみ集まる用言

例) 躓く/つまずく

Iに関しては，一意用例パターンが全く集まっておらず，曖昧用例パターンからの補完が必須である．II-iに関しても一意用例パターンの数が少なく，よりカバレッジの高い格フレーム辞書を構築するには，曖昧用例パターンから補う必要がある．

II-i, II-iiの分類については，一意用例パターンが曖昧用例パターンの用例数より少なく，かつ，かな表記で用いられやすいと人手で判断されたものをII-iとし，その他をII-iiとした．

[1]これは「ガン」という語の候補代表表記で,「ガン/がん」は銃の意である．

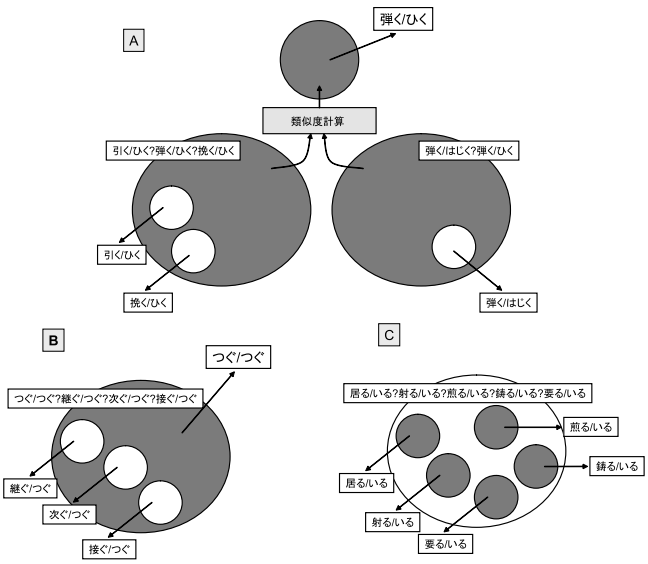

図 2: 用例パターンの振り分け方

用例パターンの振り分け

曖昧用例パターンの用言の曖昧性を解消し，用言の代表表記を決定して，一意用例パターンとすることを，「用例パターンの振り分け」と呼ぶ．用例パターンの振り分けについては，用言の種類に応じて，以下の 3 通りの方法を用いる．

A (I-i)
補う候補を含む曖昧用例パターンが漢字表記由来のものとかな表記由来のものの二種類ある場合，候補以外の用言の一意用例パターンと類似した用例パターンをそれぞれから除き，残った二種類の用例パターンの中から類似したものを目的の用言の一意用例パターンとする．また除いたものは類似度の高かった用言の一意用例パターンとする．

B (I-ii, II-i)
補う候補を含む曖昧用例パターンが一種類の場合，候補以外の用言の一意用例パターンと類似した用例パターンを除く．残った曖昧用例パターンを全て用例の不十分な用言の一意用例パターンとする．

C (II-ii)
B とは異なり，残った曖昧用例パターンを振り分けず，捨てる．

これらの振り分け手法を模式的にあらわしたものが図 2 である．

類似度は直前格，直前格要素を用いて計算する．直前格が共通のものは直前格要素間の類似度を用い，直前格が共通でないものは 0 とする．ただし，＜補文＞，＜数量＞といったものが直前格要素であるものは複数の用言の用例が混ざっている場合が多いため，用例の振り分けの対象外とした．さらに誤変換等による誤字を考慮して，高頻度の用例パターンと低頻度の用例パターンとの類似度は直前格要素に関わらず 0 とした．

5 曖昧性解消実験

前節までで述べた手法で代表表記化されたシソーラス，格フレーム辞書の有効性を示すために，これらを用いた格解析を行い，かな表記語の曖昧性解消実験を行った．

格解析におけるかな表記語の曖昧性解消は [3] の手法を用いた．用言の曖昧性解消は，選択した格フレームの用言の代表表記に決定することで行い，名詞の曖昧性解消は，格フレーム中の格要素との類似度が最高となる代表表記に決定することで行った．

本手法で代表表記化されたシソーラスと格フレーム辞書を用いた解析と，従来の解析とを比較すると，以下のような解析誤りを防ぐことができた．○が本手法で選ばれた正しい解析結果であり，×が従来の誤った解析結果である．また，無印はその他の候補を表す．

○有る/ある　×合う/あう　会う/あう
実力の裏付けが【あって】のことだった。

○癌/がん　×ガン/がん　雁/がん
【ガン】の予防に効果が期待できる。

○弾く/ひく　弾く/はじく
ギターを【弾いた】経験がなくても大丈夫。

6 おわりに

本研究では，主な自然言語リソースであるシソーラスと格フレーム辞書の代表表記化を行った．また，それらを用いた格解析において，本手法の有効性を確認した．

今後の課題としては，名詞格フレームなどのその他の自然言語リソースの代表表記化が挙げられる．また，代表表記のメンテナンスが行われた際にこれらのリソースも再構築が必要となるため，再構築にかかるコストの低減を検討したい．

参考文献

[1] 黒橋禎夫, “言語のセマンティックス”, 人工知能学会誌, Vol.21, No.6, pp.718-723 (2006.11)

[2] 河原大輔, 黒橋禎夫, “格フレーム辞書の漸次的自動構築”, 自然言語処理, Vol.12, No.2, pp.109-131 (2005)

[3] 岡部浩司, 河原大輔, 黒橋禎夫, “格フレームを用いたかな表記語の曖昧性解消”, 言語処理学会第 12 回年次大会, pp.1115-1118 (2006)