修 士 論 文

代表表記による
自然言語リソースの整備

指導教員　　江崎 浩 教授

東京大学大学院
情報理工学系研究科 電子情報学専攻

学籍番号・氏名　56410　岡部 浩司

提 出 日　　平成19年2月2日

目 次

概要

日本語には送り仮名の違いや漢字表記とかな表記の違いなど，様々な表記揺れが存在する．計算機がこれらの表記揺れを同じ語だと理解するために，形態素解析器 JUMAN では代表表記によって表記揺れをまとめている．しかし，構文解析で用いるシソーラスや格フレームといった自然言語リソースには表記揺れが混在して記載されており，そのまま用いて解析を行うと表記揺れによって適切なマッチングが行えず，正しい解析を行えないという問題があった．本研究では，それらの自然言語リソースを代表表記を用いて整備し，リソース中の表記揺れの解消を行った．

第1章　序論

日本語の言語処理の抱える問題の一つとして,「蜜柑」,「みかん」,「ミカン」といった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．日本語は漢字，ひらがな，カタカナと多くの文字を持ち，同じ単語が漢字表記とかな表記の両方で書かれることや，送り仮名に複数の付け方があることなどがあり，同じ語に対して複数の表記の仕方がある．諸外国語でも表記揺れは存在するが，日本語の表記揺れは非常に変化に富んでおり，計算機での扱いが非常に困難である [5]．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．しかし，表記揺れについては同一の語として一つにまとめ，かな表記語は表すものごとに区別して扱った方が，より適切な言語処理を行えることは明らかである．そのように扱うには，どういった語が表記揺れであり，かな表記語が表すものにはどのような候補があるのかを計算機に理解させる必要がある．

形態素解析器 JUMAN5.0 では，計算機に言語を理解させるという観点から，基本語彙の選定と代表表記の付与が行われた [1]．ここでは人手で十分メンテナンス可能なサイズに辞書の語彙を制限し，それらの語彙に対して同じ語の表記揺れには，同じ代表表記を与えるということを行った．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で，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．つまり形態素解析によって表記揺れを吸収できるようになったといえる．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．そこで，これらの自然言語リソースを代表表記で扱うようにすることが求められていた．

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．さらに代表表記化されたこれらの自然言語リソースを用い，格解析の中でかな表記語の曖昧性解消を行う手法の実装を行う．

本論文は全6章から成り，その構成は以下の通りである．第2章では，JUMANに実装された代表表記について説明する．第3章では，シソーラスの代表表記化，第4章では，格フレーム辞書の代表表記化について述べる．第5章では，これらの自然言語リソースを用いたかな表記語の曖昧性解消の手法について述べ，その実験結果に対する考察を行う．最後に結論と今後の課題を第6章でまとめる．

第2章　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表される．例えば「蛍」,「ほたる」,「ホタル」の代表表記は全て同一であり,「蛍/ほたる」のように表現する．

代表表記は原則として新聞記事・ウェブ等での高頻度表記としているが，この選択に強いこだわりはなく，妥当性を主張するものでもない．重要な事は，同じ語の表記集合がまとめられ，これを通してテキストマッチングなどが適切に行われることである．ただし，IDを数字などにすることは人間にとって管理しやすいものではないので，高頻度の表記（とその読みのペア）をIDとしている．

形態素解析器JUMAN5.0[6]では代表表記を意味情報として出力でき，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．

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．

漢字とかな	拳銃 けん銃 拳じゅう けんじゅう	→	拳銃/けんじゅう
	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
	行う 行なう	→	行う/おこなう
漢字表記	色取る 彩る	→	彩る/いろどる
	奇跡 奇蹟	→	奇跡/きせき
カタカナ語	コンピュータ コンピューター	→	コンピューター/こんぴゅーたー
	デジタル ディジタル	→	デジタル/ でじたる
音便	さようなら さよなら	→	さようなら/さようなら

また,「蛍」と「ホタル」は同一のものを指す表記バリエーションであり，同じ代表表記「蛍/ほたる」を与えられるが,「円陣」と「エンジン」は異なるものを指す語のため，それぞれ「円陣/えんじん」と「エンジン/えんじん」という異なる代表表記が与えられるよう考慮されている（図2.1上部）．

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」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」

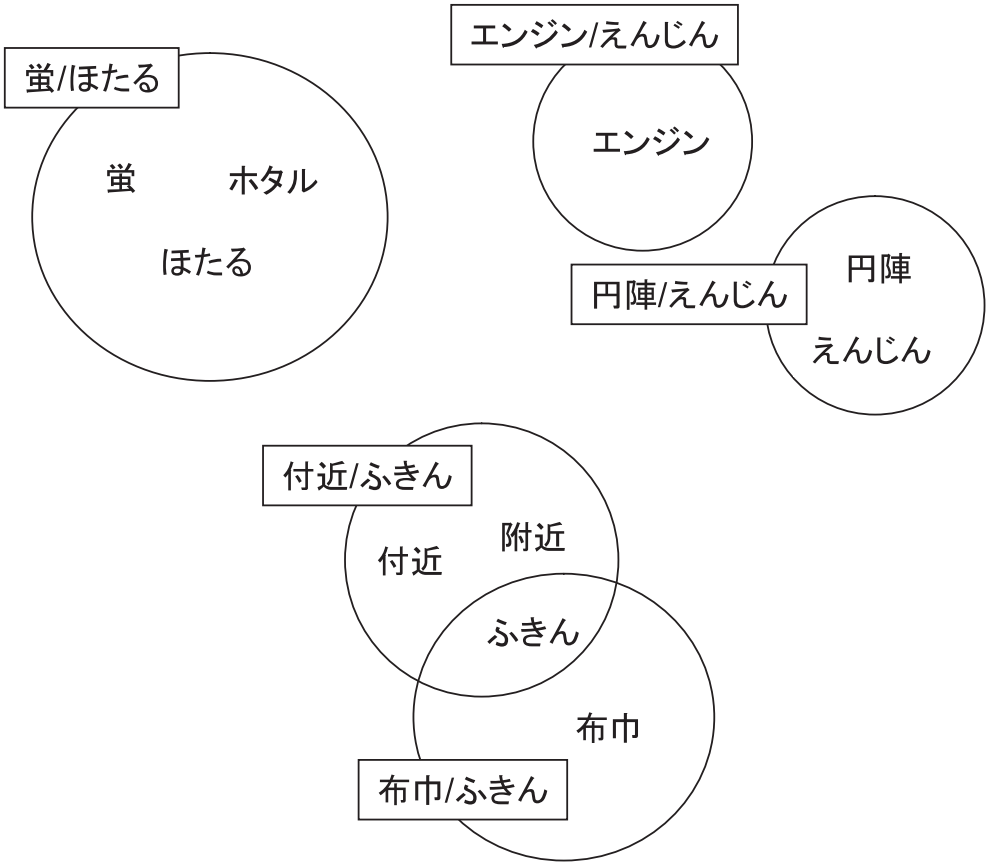

図 2.1: 代表表記によってまとめられる表記揺れ

は同一の語の表記揺れとして一つの代表表記でマージされ,「賦金」と「斧斤」は日常使用されないため辞書から削除されており,「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（図 2.1 下部）．その複数の代表表記の中から一つの代表表記を決定することで，かな表記語による曖昧性を解消することができる．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ための準備が整ったといえる．

図 2.2 は「かぜでおくれた」という文を JUMAN で形態素解析した結果である．「かぜ」,「おくれた」は共にかな表記による曖昧性を持ち，その候補としてそれぞれ「風/かぜ」と「風邪/かぜ」,「送れる/おくれる」と「遅れる/おくれる」が挙げられている．この中から「風邪/かぜ」,「遅れる/おくれる」を選択することで，かな表記語の曖昧性解消を行うことが可能である．

また，動詞，形容詞の代表表記は原形で辞書に記載されている．活用形や形容詞語幹の場合はそのまま記載されてはいない．例えば「待て」,「温厚」の代表表記はそれぞれ,「待つ/まつ」,「温厚だ/おんこうだ」である．

地名，人名，組織名，固有名詞，数詞，連語の一部は，JUMAN の辞書に代表表記が登録されていない．代表表記で単語を扱う際に，これらの語を表わす表記が

```
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 2.2: JUMAN の解析例

必要であるため，構文解析器 KNP[7] ではこれらの語に対して，疑似代表表記を与える．疑似代表表記は JUMAN の出力の見出しと読み（いずれも原形）を用いて，「見出し/読み」の形で与えられる．例えば，「アメリカ」の疑似代表表記は「アメリカ/あめりか」である．

第3章　シソーラスの代表表記化

シソーラスとは，単語の上位/下位関係，部分/全体関係，同義関係，類義関係などによって単語を分類し，体系づけた辞書である．図 3.1 のような木構造をしており，類似した語同士が近くにまとまっている．そのため，単語間の類似度を計算する際に用いられる．

日本語のシソーラスには，日本語語彙大系 [8]，分類語彙表 [9] などがある．英語のものでは，ロジェのシソーラス [10]，WordNet[11] などが挙げられる．本研究で代表表記化を行ったシソーラスは分類語彙表であり，その一部を図 3.2 に示す．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，分類語彙表には次のように「ミステリー」という語のエントリーがある．

11030050104 ミステリー みすてりい

しかし,「ミステリ」という語のエントリーはなく，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．シソーラスが代表表記で記載されていれば,「ミステリー」と「ミステリ」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，

11030050104 ミステリー/みすてりー

というエントリーから意味コードを得ることができる．

また，分類語彙表には，下記のような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．

15050040702 あく あく

このエントリーは「灰汁」の意味で記載されているのだが，見出し語がかな表記されているため,「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．

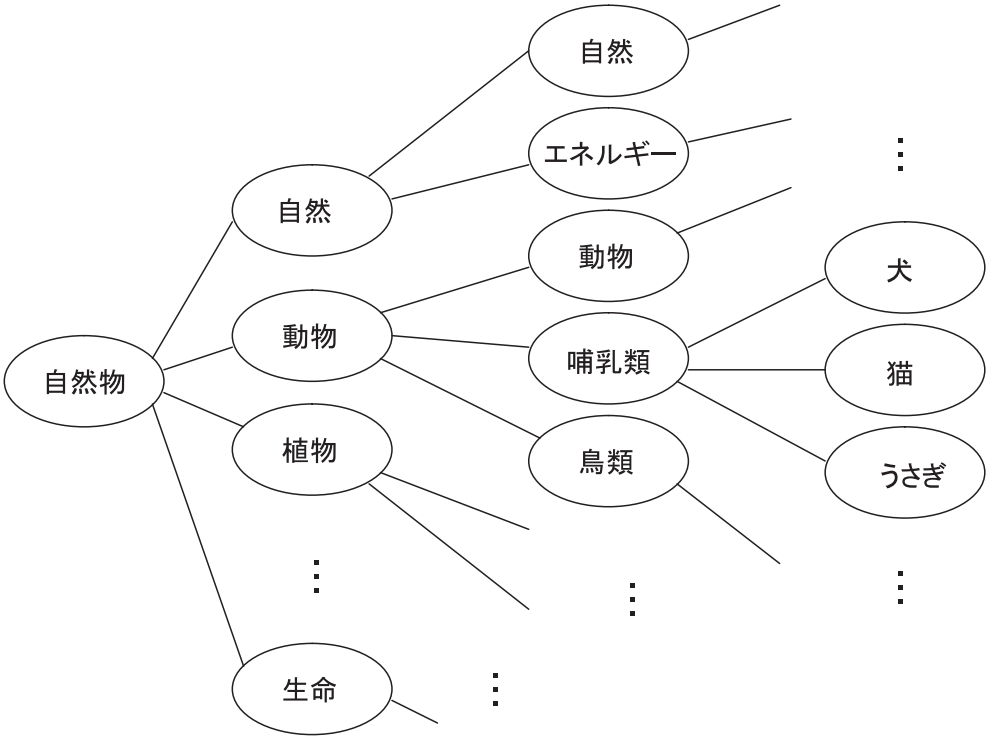

図 3.1: シソーラス

```
11030030104 虚偽 きょぎ
11030030201 夢物語 ゆめものがたり
11030030202 一場の夢 いちじょうのゆめ
11030030203 空中楼閣 くうちゅうろうかく
11030040101 仮性 かせい
11030040102 真性 しんせい
11030040201 仮想現実 かそうげんじつ
11030040202 バーチャルリアリティー ばあちゃるりありてぃい
11030040301 試し ためし
11030050101 不思議 ふしぎ
11030050102 七不思議 ななふしぎ
11030050103 神秘 しんぴ
11030050104 ミステリー みすてりい
11030050105 オカルト おかると
11030050201 複雑怪奇 ふくざつかいき
```

図 3.2: 分類語彙表

3.1 ルールによる自動的な代表表記変換

はじめに，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．

1.a JUMAN の辞書に一致する語が存在すれば，その語の代表表記を出力する．また,「飲料水」など「〜水」と用いるようなものは，分類語彙表の見出しが「−水」となっているが，これは「水」とみなす．

(1.a) 接続 せつぞく → 接続/せつぞく
持ち味 もちあじ → 持ち味/もちあじ
奇蹟 きせき → 奇跡/きせき
−水 すい → 水/すい

1.b JUMAN の辞書に一致する語が複数存在すれば，?で代表表記を繋げて全て列挙する．ただし,「堪忍/かんにん」など，同じ代表表記で名詞と感動詞など複数の品詞を持つようなものは，曖昧性があるとは考えずに一つだけ出力する．

(1.b) そば そば → 傍/そば?蕎麦/そば
くし くし → 串/くし?櫛/くし?駆使/くし
堪忍 かんにん → 堪忍/かんにん

1.c ナ形容詞の語幹のために，表記と読みの語尾に「だ」を付けた語が形容詞として JUMAN の辞書にあればその代表表記を全て列挙する．

(1.c) 不滅 ふめつ → 不滅だ/ふめつだ
平和 へいわ → 平和だ/平和だ
でたらめ でたらめ → でたらめだ/でたらめだ

1.d カタカナ語の長音の読みは，分類語彙表では母音で記載されており JUMAN の辞書の表記と異なるため，読みの表記は見出しをひらがな表記に置換したものを用いる．

(1.d) アイテム あいてむ → アイテム/あいてむ
ショート しょおと → ショート/しょーと
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．(KNP を用いるのは，品詞

変更した語の代表表記を得るためである.）この時，入力は文ではなく，単語であるため，JUMAN の解析ルールに命令形，終助詞終わりを許さないルールを用いる．

2.a 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．この時，1.b，1.c，1.d と同様の処理を行う．読みが一致した時，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．

(2.a) おたふく風邪 おたふくかぜ → おたふく/おたふく+風邪/かぜ
粒あん つぶあん → 粒/つぶ+暗/あん?案/あん?餡/ あん
安全装置 あんぜんそうち → 安全だ/あんぜんだ+装置/そうち
スピードアップする すぴいどあっぷする
→ スピード/すぴーど+アップ/あっぷ+する/する

2.b 品詞変更した動詞だと解析されれば，品詞変更した形の代表表記を出力する．品詞変更した動詞の代表表記には末尾に v が付く．

(2.b) 動き うごき → 動き/うごき v
放射能漏れ ほうしゃのうもれ → 放射能/ほうしゃのう+漏れ/もれ v
育て親 そだておや → 育て/そだて v+親/おや

2.c 複合語には濁音化しているものがあるため，代表表記の読みの一文字目を濁音化し，読みが一致するかどうかをチェックする．

(2.c) 掘りごたつ ほりごたつ → 掘り/ほり v+炬燵/こたつ
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

2.d 代表表記を持たない語（地名，人名，組織名，固有名詞，数詞，連語の一部）と解析された場合は，JUMAN の出力した見出しと読みを用いて，「見出し/読み」の形で KNP から出力される疑似代表表記を用いる．

(2.d) アメリカ あめりか → アメリカ/あめりか
ノーベル賞 のおべるしょう → ノーベル/のーべる+賞/しょう
世界一 せかいいち → 世界/せかい+一/いち

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，JUMAN の辞書に採用されていない，日常使用しない語だと判断し，代表表記化した分類語

彙表のエントリーから除く．

(3) 抜糸 ばっし → NO(抜糸 ばっし)
殷鑑 いんかん → NO(殷鑑 いんかん)
ジェネレーション じぇねれえしょん
→ NO(ジェネレーション じぇねれーしょん)

以上のルールを用いて，分類語彙表の各エントリーの見出しと読みのペアを代表表記列に変換した．

3.2 人手による正しい代表表記の判断

3.1 において，分類語彙表の各エントリーをルールにより自動的に代表表記に変換したが，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?を含み複数の候補が列挙され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) 12410280204 ねぎ ねぎ → 葱/ねぎ
(ii) 14310310202 粒あん つぶあん → 粒/つぶ+暗/あん?案/あん?餡/あん
(iii) 11130040102 もどき もどき → 喪/も?藻/も+土器/どき

(i) のエントリーは「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．(iii) では,「もどき」を JUMAN の辞書に採用しておらず，KNP の解析によって誤った複合語が候補に挙げられている．

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し，その代表表記に決定するか，その語は日常使用しないので 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をしてしまうことを防ぐことができる．

```
11030030104 虚偽/きょぎ
11030030201 夢/ゆめ+物語/ものがたり
11030030203 空中/くうちゅう+楼閣/ろうかく
11030040102 真性/しんせい
11030040201 仮想/かそう+現実/げんじつ
11030040301 試し/ためし v
11030050101 不思議だ/ふしぎだ
11030050102 七不思議/ななふしぎ
11030050103 神秘だ/しんぴだ
11030050104 ミステリー/みすてりー
11030050105 オカルト/おかると
11030050201 複雑怪奇だ/ふくざつかいきだ
```

図 3.3: 代表表記化分類語彙表

最後に分類語彙表の連続するエントリーに同じ代表表記が与えられた場合には，そのうちの一つだけを残すようにする

11962680201	わん わん		椀/わん
11962680202	椀 わん	→	—
11962680203	碗 わん		—

以上によって代表表記化された分類語彙表の一部を図 3.3 に示す.

第4章　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文章中の表記の原形を用いて構築されており，表記揺れがそれぞれ異った表記で格フレーム辞書中に存在していた．

そこで，シソーラスと同様に格フレーム辞書の代表表記を行った．代表表記化することで，今まで「表す」と「表わす」と別々の用言として存在していた表記揺れの格フレームを「表す/あらわす」の格フレームとして一つにまとめることができる．同様に格要素における表記揺れもまとめることができ，例えば「積む/つむ」の格フレームにおけるヲ格の格要素の「研鑽」と「研さん」をまとめることができる．

また，格フレーム辞書を構築する中で,「あめ」という「雨」,「飴」のどちらの意味か分からない格要素の表記を，その他の格要素との類似度を考慮することで,「雨/あめ」や「飴/あめ」と直し，格要素の持つ曖昧性の解消も行う．

さらに,「掻く/かく」といったかな表記で用いられる頻度が高く，用例が十分に集まらない用言に対して,「かく」といった「欠く/かく」,「書く/かく」,「掻く/かく」の曖昧性を持った用言から適切な用例を見つけ出し，用例の補完を行う．

4.1　格フレーム辞書構築

4.1.1　格フレーム辞書構築の手順

ここでは，代表表記を用いた格フレーム辞書の構築方法を示す．格フレーム辞書の構築方法は河原 [2] の手法を用い，用言と格要素を代表表記で収集するようにした．

格フレーム辞書を大規模コーパスから自動的に構築するためには，まず，構文解析をしなければならないが，ここで解析誤りが問題となる．この問題は有る程度信頼度の高い係り受けだけを学習に用いることで対処することができる．むしろ問題なのが用言の用法の多様性で，格フレームは用言の用法ごとに作成する必要がある．これに対処するために，用言と直前の格要素を単位として用例を収集し，クラスタリングを行う．

まず，格フレームに関する語句の定義について述べる．

用例 テキストコーパス中に存在する述語項構造を，用言と格要素の組で表したもの．

例)（格の末尾に*が付くものが直前格である．）
文：以来、４人でチームを組み、毎週、特訓を重ねている。
↓
用例：
組む/くむ:動　＜数量＞人/にん:４/よん:デ格　チーム/ちーむ:ヲ格*
重ねる/かさねる:動　＜時間＞:毎週/まいしゅう:時間　特訓/とっくん:ヲ格*

用例パターン 用言と直前格要素が共通するもの同士で用例をまとめたもの．

例)　（代表表記の末尾の数字は頻度を表わす．）
組む/くむ:動 12
＜ガ格＞　＜数量＞人/り:5, 選手/せんしゅ:5,…
*＜ヲ格＞　ペア/ぺあ:1203
＜ニ格＞　ダム/だむ:2, レース/れーす:1, 監督/かんとく:1,…
＜ト格＞　＜補文＞:2, 選手/せんしゅ:1,…
＜デ格＞　＜数量＞:12, ＜数量＞人/り:11, 人/じん?人/ひと:5, 車/くるま?車/しゃ:3, 男女/だんじょ:3

格フレーム 用例パターンのクラスタリングを行い，用例パターンをまとめたもの．

例)　組む/くむ:動 3
＜ガ格＞　選手/せんしゅ:8, 刑事/けいじ:8, 人/ひと:4,…
＜*ヲ格＞　コンビ/こんび:2629, ペア/ぺあ:1203, グループ/ぐるーぷ:511,…
＜ニ格＞　スタジオ/すたじお:6, 実際/じっさい:5, 後/あと?後/のち:3,…
＜ト格＞　誰/だれ:3, 人/ひと:1, ベテラン/べてらん:1,…
＜デ格＞　人/ひと:6, 子供/こども:4, 仕事/しごと:4,…

次に代表表記を用いた格フレーム辞書の構築手順を以下に示す．

1. テキストコーパスの各テキストを KNP を用いて構文解析し，その結果から信頼度の高い述語項構造のみを抽出し，用例を得る

   1.a ここで収集される用例の用言，格要素の表記を KNP の解析結果から得られる代表表記とする．複合語は”+” でそれぞれの代表表記を連結し，品詞変更した語の代表表記には末尾に v が付く表記とする．

   1.b 複数の可能性が考えられ曖昧性があるものは，それぞれの候補の代表表記列を?で連結した形で表記する．

ここまでで得られる用例の例を以下に示す．

例) 地方が驚くべきはやさで過疎化している。
過疎/かそ+化/か:動 早い/はやい+さ/さ?速い/はやい+さ/さ:デ格 地方/ちほう:ガ格*

2. 次に，用例を用言と直前の格要素の組ごとにまとめ，用例パターンを得る．この時，直前格の閾値頻度を 5 とし，これより少ない頻度の直前格を持つ用例パターンは用いないこととした．

3. あらゆる二つ組の用例パターンの類似度を計算し，直前の格要素の意味コードの固定を行う．その中で曖昧性を持つ直前の格要素の代表表記を決定する．意味コードの固定，および類似度の定義については後述する．その後，用例パターン間の類似度が閾値を越える組について，用例パターンのマージを行う．

4. 次に直前格を限定しない用例パターンのクラスタリングを行う．その後，直前格以外の格要素で曖昧性を持つものの意味コードの削除を行う．意味コードの削除の定義についても後述する．こうして，最終的な格フレームを得る．

また，格要素の半数以上の用例が分類語彙表の主体のカテゴリーに含まれる格には，＜主体＞という意味マーカを付与する．これを持つ格は人，組織といった主体的要素をとることを示す．その他にも＜数量＞，＜補文＞といった意味マーカがある．

4.1.2 用例パターン間の類似度

用例パターンのクラスタリングに用いる用例パターン間の類似度は，格の一致度と用例群間の類似度の積とする．

まず，単語 (代表表記) e_1, e_2 間の類似度 $sim(e_1, e_2)$ を代表表記化した分類語彙表を利用して，以下のように定義する．

$$sim(e_1, e_2) = max_{x \in s_1, y \in s_2} sim(x, y) \tag{4.1}$$

$$sim(x, y) = \frac{2L}{l_x + l_y} \tag{4.2}$$

ここで，x, y は意味コードであり，s_1, s_2 はそれぞれ e_1, e_2 の分類語彙表における意味コードの集合である．$sim(x, y)$ は意味コード間 x, y 間の類似度であり，l_x, l_y は，x, y のシソーラスの根からの階層の深さ，L は x と y の意味属性で一致している階層の深さを表す．この類似度 $sim(x, y)$ は 0 から 1 の値をとる．

用例パターン P_1, P_2 の格の一致度 $cs(P_1, P_2)$ は，P_1, P_2 のそれぞれに含まれる全ての格要素に対する，P_1, P_2 の共通格に含まれている格要素の割合を求め，それらの積の平方根とする．計算式は次のように定義する．

$$cs(P_1, P_2) \quad = \quad \sqrt{\frac{\sum_{i=1}^{n} |E_{1cc_i}|}{\sum_{i=1}^{l} |E_{1c1_i}|} \times \frac{\sum_{i=1}^{n} |E_{2cc_i}|}{\sum_{i=1}^{m} |E_{2c2_i}|}} \tag{4.3}$$

ただし，用例パターン P_1 中の格を $c1_1, c1_2, \cdots, c1_l$，用例パターン P_2 中の格を $c2_1, c2_2, \cdots, c2_m$，P_1 と P_2 間の共通格を $cc_1, cc_2, \cdots, cc_n$ とする．また，E_{1cc_i} は P_1 内の格 cc_i に含まれる格要素群であり，$E_{2cc_i}, E_{1c1_i}, E_{2c2_i}$ も同様である．$|E_{1cc_i}|$ などの絶対値記号は頻度を表す．

用例パターン P_1, P_2 の共通格 cc_i に含まれる格要素群間の類似度 sim_{cc_i} は，それぞれの格要素ごとに最も類似度の高い格要素を相手の格要素群からみつけ，その類似度の平均とし，

$$sim_{cc_i} = \frac{\sum_{e_1 \in E_{1cc_i}} |e_1| max_{e_2 \in E_{2cc_i}} sim(e_1, e_2) + \sum_{e_2 \in E_{2cc_i}} |e_2| max_{e_1 \in E_{1cc_i}} sim(e_1, e_2)}{|E_{1cc_i}| + |E_{2cc_i}|} \tag{4.4}$$

と定義する．頻出する格は重要度が高いと考えて，各共通格の格要素群間の類似度に頻度重みを付け，P_1, P_2 間の格要素群間の類似度 $sim_E(P_1, P_2)$ を以下のように定義する．

$$sim_E(P_1, P_2) = \frac{\sum_{i=1}^{n} \sqrt{|E_{1cc_i}|}\sqrt{|E_{2cc_i}|} sim_{cc_i}}{\sum_{i=1}^{n} \sqrt{|E_{1cc_i}|}\sqrt{|E_{2cc_i}|}} \tag{4.5}$$

用例パターン P_1, P_2 間の類似度は，格の一致度 $cs(P_1, P_2)$ と，格要素群間の類似度 $sim_E(P_1, P_2)$ の積とし，次のようにして計算する．

$$\text{類似度} \quad = \quad cs(P_1, P_2) \cdot sim_E(P_1, P_2) \tag{4.6}$$

この定義を用いて，類似度を求めた具体的な例を図 4.1 に示す．

4.2 曖昧性を持つ格要素の処理

格要素を代表表記化することで，表記揺れが一つの代表表記でまとめられて扱われるようになった．しかし，かな表記などで曖昧性のある格要素は,「ガン/がん?癌/がん?雁/がん」（「ガン」の表記の曖昧性）のように複数の表記が候補に挙げられている．これらの格要素は本来の意味以外の意味コードも持っているため，格フレームを用いた解析時の誤り原因となることがある．

したがって，格フレーム構築の際に，曖昧性を持つ格要素に対して曖昧性を解消する処理を行った．直前格要素に対しては「意味コードの固定」，その他の格要素に対しては「意味コードの削除」を行った．

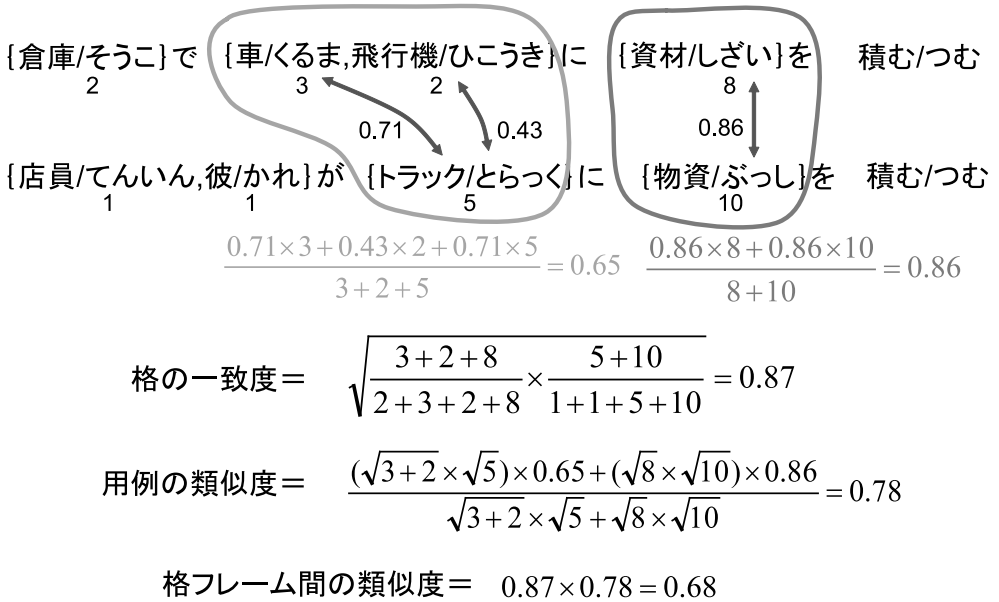

図 4.1: 格フレーム間の類似度

4.2.1 意味コードの固定

意味コードとはシソーラスに記載されている各語のコードを指す．多義である語の多義性を考慮せずに単純にクラスタリングを行うと，全く別の意味コードの語が同一の用例パターンにマージされてしまう．この問題に対処するために，直前格の格要素の意味コードを固定する処理を行う．

また，意味コードを固定することによって，曖昧性をもったかな表記語の格要素がどの意味で用いられてるかが判別でき，曖昧性を解消することができる．

意味コードの固定は次の手順で行う．

1. 用例パターン間の類似度を全てのペアについて計算する．格フレーム間の類似度の尺度は前節で述べた用例パターン間の類似度と同じものを用いる．

2. 類似度の高い用例パターンのペアの直前格の格要素 (p, q) から順に，その意味コードを類似度を計算する時に用いたものに固定する．その後，p, q を含む用例パターンに関係するペアの類似度を再計算する．この時，p, q が代表表記に複数の候補を持ち，?で代表表記が繋げられている語の場合，固定された意味コードを持つものに代表表記を変更する．

3. 類似度が閾値より小さくなるまで順に意味コードの固定と類似度の再計算を繰り返す．

以下に，この処理の例を示す．用言「飛ぶ/とぶ」の直前格が「ガン/がん?癌/がん?雁/がん」,「雁/がん」,「鳥/とり」,「ミサイル/みさいる」であり，類似度が閾値

（0.6 とする）を越えるペアは以下の４通りであったとする.（＜＞内は類似度を計算する時に用いられた意味コードの属するカテゴリー名である．右端の数字は用例パターン間の類似度を示す.）

(1) ガン/がん?癌/がん?雁/がん＜鳥類＞ 雁/がん＜鳥類＞ 0.926
(2) ガン/がん?癌/がん?雁/がん＜鳥類＞ 鳥/とり＜鳥類＞ 0.714
(3) ガン/がん?癌/がん?雁/がん＜武器＞ ミサイル/みさいる＜武器＞ 0.678
(4) 鳥/とり＜鳥類＞ 雁/がん＜鳥類＞ 0.668

ここで「ガン/がん?癌/がん?雁/がん」は＜武器＞, ＜鳥類＞, ＜病気・体調＞の三つのカテゴリーの意味コードを持つ．まず，最も類似度の高い用例パターンの組は (1) であり，この類似度を取る時の「ガン/がん?癌/がん?雁/がん」の意味コードのカテゴリーは＜鳥類＞である．したがって，ここで「飛ぶ/とぶ」の直前格の全ての「ガン/がん?癌/がん?雁/がん」の意味コードを＜鳥類＞に属する意味コードに固定する．そして,「ガン/がん?癌/ がん?雁/がん」を含むペアの (2), (3) の類似度を再計算する．その結果，(2) の類似度は変わらないが，(3) は，

(3) ガン/がん?癌/がん?雁/がん＜鳥類＞ ミサイル/みさいる＜武器＞ 0.261

となり，閾値以下の類似度となるため,「ミサイル/みさいる」の用例パターンはクラスタリングされない．さらに,「ガン/がん?癌/がん?雁/がん」の三つの代表表記のうち，＜鳥類＞に含まれる意味コードを持つものは「雁/がん」であるから,「ガン/がん?癌/がん?雁/がん」の表記を「雁/がん」に変更する．

4.2.2 意味コードの削除

意味コードの削除は，ある格フレームの格の中から他の用例との類似度の低い意味コードをみつけるもので，これを用いて，クラスタリングの後に残っている直前格以外の格要素の曖昧性を緩和することができる．意味コードの削除はクラスタリングされた格フレームの格ごとに行う．意味コード削除の手順を以下に示す.

1. 目的の格の格要素とその頻度のペアの集合を，意味コードとその頻度のペアの集合に変換する．
2. 一つの意味コードに対して，各意味コード（自分自身とも）との類似度を計算する．
3. スコアの高いものから順に頻度が全体の 1/5 となるまで，頻度重みを付けて類似度の和をとり，それをその意味コードのスコアとする.（頻度重みは頻度積，自分自身とは (頻度) ∗ (頻度 – 1)）

4. 計算されたスコアが閾値以下の時，その意味コードは同じ格の他の格要素と類似度が低いので，削除対象として列挙する．スコアの閾値の値は 0.4 とする．

次に，意味コード削除を用いた格要素の曖昧性の緩和の手順を述べる．

1. ?でつながれた曖昧性をもつ格要素を?で分割し，候補代表表記を挙げる．
2. ある候補代表表記の持つ意味コードが全て意味素削除されていれば，その候補代表表記ではないと判断し，除外する．
3. 2 で残った代表表記を?でつないだ表記を新たな表記とする．2 で候補代表表記が全て除外されていたら何も行わない．

以下に意味コードの削除の処理を具体例を挙げて示す．次のような「焼く/やく」の格フレームがあったとする．

焼く/やく:動
＜デ格＞　オーブン/おーぶん:18　釜/かま?鎌/かま?窯/かま:2

まず，デ格の格要素と頻度のペアを意味コードと頻度のペアに直すと次のようになる.（意味コードを分かりやすく＜表記＞で表している.）

＜オーブン＞:18　＜釜 1＞:2　＜釜 2＞:2　＜鎌＞:2　＜窯＞:2

＜鎌＞という意味コードに注目してみると，一番類似度が高いものは自分自身の＜鎌＞で，類似度は 1.0 である．次は＜釜 1＞との類似度で 0.43 であり，次が＜オーブン＞との類似度の 0.29 である．総頻度数は 26 であるから，オーブンの時点で類似度を取った意味コードの頻度が 2+2+18=22 で 1/5 を超える．スコアは，

$$\text{スコア} = \{(2-1) * 1.0 + 2 * 0.43 + 18 * 0.29\}/22 = 0.32$$

となり閾値を下回るので＜鎌＞の意味コードは削除される．同様に，＜釜 1＞，＜釜 2＞の意味コードも削除され,「鎌/かま」と「釜/かま」の意味コードは全て削除されるため,「釜/かま?鎌/かま?窯/かま」の表記は「窯/かま」に変更される．

仮に，＜釜 1＞は削除されなかったとしたら「釜/かま」の意味コードはまだ残っているため,「釜/かま?窯/かま」へと変更される．

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．

本節では，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．

4.3.1 用言の分類

まず，用言表記が曖昧性を持つか，持たないかで用例パターンを次のように分類する．

一意用例パターン 曖昧性をもたない用言の用例パターン．
例）掻く/かく:動
弾く/はじく:動

曖昧用例パターン 曖昧性をもつ用言の用例パターン．
例）欠く/かく?書く/かく?掻く/かく:動
弾く/はじく?弾く/ひく:動

曖昧用例パターンの用言表記は,「欠く/かく?書く/かく?掻く/かく」のように，候補となる代表表記を列挙された形で表されている．

さらに用例パターンが，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（引く/ひく?弾く/ひく?挽く/ひく，弾く/はじく?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?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（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ある程度集まる用言
例) 会う/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，より高品質な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-iiは用例パターンを補うことが可能だが，すでに十分な用例パターンが集まっており，必要性は高くない．また，IIIは曖昧用例パターンがなく，すでに適切に表記揺れが一つの用言としてまとめられており，補完をする必要はない．

4.3.2 用例パターンの振り分け手法

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，一意用例パターンに含めることを,「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以下の3通りの方法が考えられる．

A 補う候補を含む曖昧用例パターンが漢字表記とかな表記の二種類ある場合，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．

B 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目的の用言の一意用例パターンとする．

C Bとは異なり，残った曖昧用例パターンを振り分けず，曖昧用例パターンのままとする．

これらの振り分け手法を模式的にあらわしたものが図4.2である．

次にどの用言をどの振り分け手法で振り分けるかについての対応について，具体例を交えながら詳細に記す．

まず，4.3.1のIで述べた，曖昧用例パターンのみしか集まらない用言をJUMANの辞書から抽出した．これは，その用言の表記がいずれも曖昧性を持ち，活用によっても曖昧性が解消されないものである．

I-i, I-iiにはそれぞれ，10種類，4種類の用言がある．

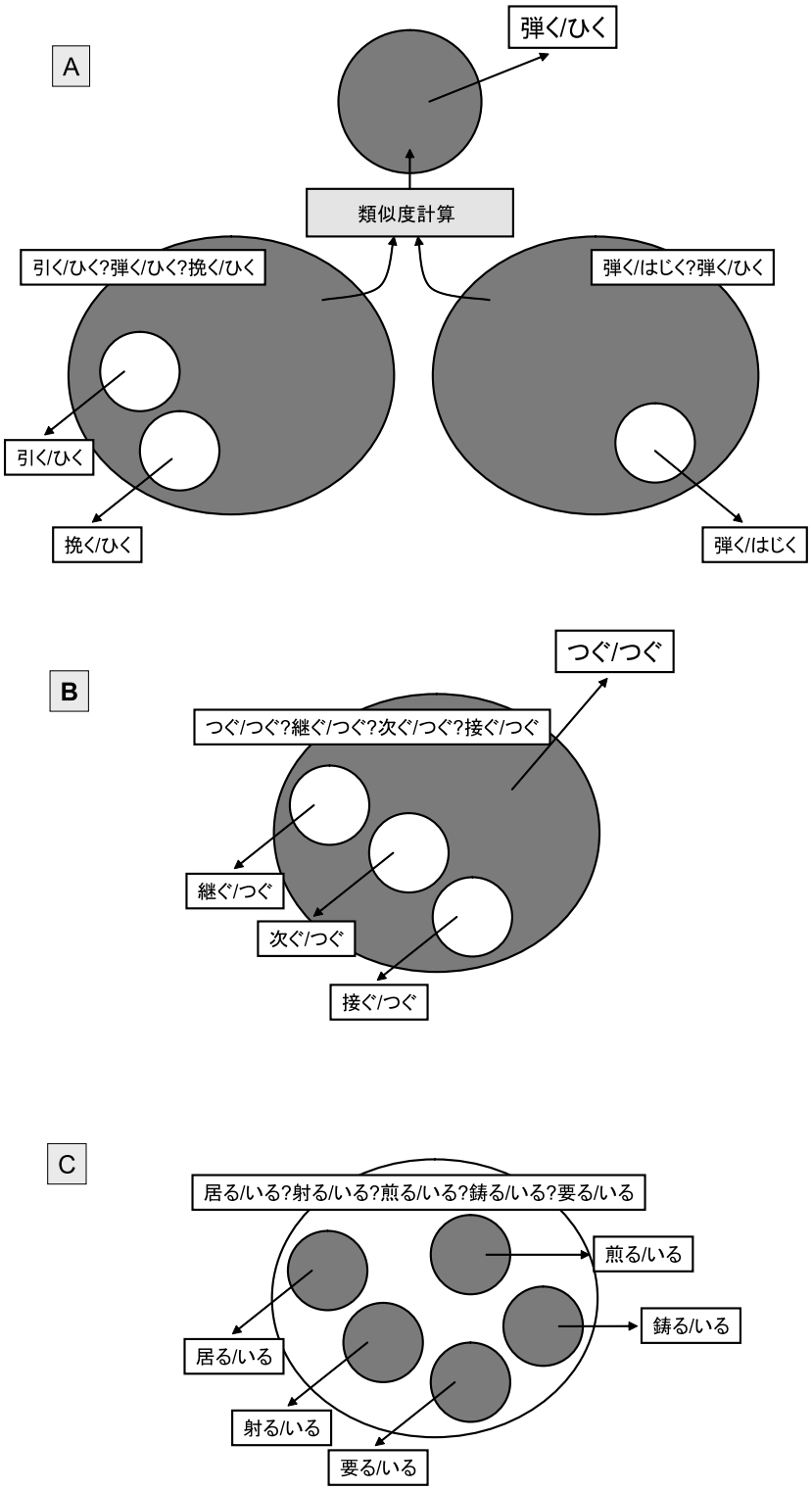
A
弾く/ひく
類似度計算
引く/ひく?弾く/ひく?挽く/ひく
弾く/はじく?弾く/ひく
引く/ひく
挽く/ひく
弾く/はじく
B
つぐ/つぐ
つぐ/つぐ?継ぐ/つぐ?次ぐ/つぐ?接ぐ/つぐ
継ぐ/つぐ
次ぐ/つぐ
接ぐ/つぐ
C
居る/いる?射る/いる?煎る/いる?鋳る/いる?要る/いる
煎る/いる
鋳る/いる
居る/いる
射る/いる
要る/いる

図 4.2: 振り分け方

I-i 漢字表記，かな表記のどちらも曖昧な用言（10 個）
弾く/ひく　摘む/つむ　留める/とめる　突く/つく　怒る/おこる
解ける/とける　開く/あく　開ける/あける　空く/あく　空く/すく

I-ii 漢字表記がなく，かな表記が曖昧な用言（4 個）
くれる/くれる　つぐ/つぐ　なまる/なまる　やめる/やめる

I-i に該当する用言の用例パターンは，例えば「弾く/ひく」であれば「弾く/はじく?弾く/ひく」や「引く/ひく?弾く/ ひく?挽く/ひく」といった，集めたい用言を候補に含む曖昧用例パターンが，漢字表記のものとかな表記のもので，二種類存在するので，A の手法で補完を行う．

I-ii に該当する用言の用例パターンは，例えば「くれる/くれる」であれば集めたい用言を候補に含む曖昧用例パターンが「くれる/くれる?暮れる/くれる」の一種類しかないため，B の手法で補完を行う．

次に II-i, II-ii の区別の基準であるが，以下の基準を満たすものを II-i とする．

基準 1 その用言の全用例数が，その用言を候補に含む曖昧用例パターンの全用例数より少ない．

基準 2 さらに，その用言は漢字表記に比べて，かな表記が多用されると人手で判断される．

例を挙げると,「千切る/ちぎる」は用例数が 28 であるのに対し,「契る/ちぎる?千切る/ちぎる」の用例数が 706 であることから基準 1 を満たす．さらに,「千切る」よりも「ちぎる」と書くことが多く基準 2 も満たすため，II-i と判断できる.「契る/ちぎる」は用例数は 13 であり基準 1 を満たすが,「契る」と書くことが多いため基準 2 を満たさず II-ii と判断される.（図 4.3 上部参照）

また,「会う」,「合う」に関してはどちらも用例数が約 5 万あり,「会う/あう?合う/あう」の用例数は 1 万程度であるため，基準 1 を満たさず，II-ii と判断される．（図 4.3 下部参照）

ここで，曖昧性のある用言の候補用言の中で一つだけが II-i に分類される時，曖昧性のある用言の用例から他の候補用言の用例と類似したものを取り除くと，残りが II-i に分類される用言の用例と考えることができる．したがって B の手法で用例を補う．

候補に II-i に分類される用言が一つもない，または複数ある時は，どれか一つの用言に残った用例を振り分けてしまうと誤った用例を振り分けてしまう可能性が高いため，C の手法で用例を振り分けるのみとする．

4.3.3 分類手法の詳細

ここでは実際の分類手法についての詳細な手順を述べる．分類は以下の手順で行う．

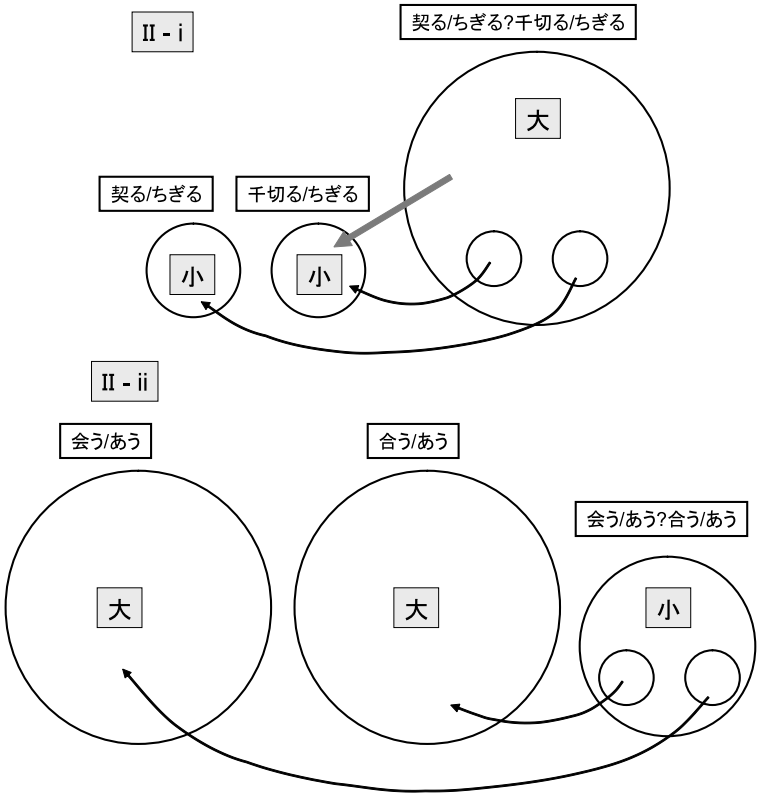

図 4.3: 用例数の大小

1. 最初に，A, B, C のいずれの場合も，集めたい用言を含む曖昧性のある用言の用例パターン群を用意する．また，以下の用例パターンは複数の用言の用例が混ざっている可能性が高く，対象外とする．

   - ＜補文＞，＜数量＞，＜時間＞といった汎化されたマーカが直前格要素であるもの
   - 直前格がヲ格，ニ格以外であり，かつ直前格要素が＜主体＞の意味コードをもつもの

2. それぞれの用例パターンについて，格候補用言の用例パターンとの類似度をとる．類似度については後述する．類似度の閾値を 0.8 とし，閾値を越え，さらに他の候補用言はその類似度未満である時，その用例パターンを類似度の最大値を取った用言に振り分ける．

3. A では 2 の後で残った複数の曖昧性のある用言から共通の直前格，直前格要素をもったものを，I-i の用言に補う．B では 2 の後に残った用例を I-ii または II-i の用言に振り分ける．

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．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．

以上により，曖昧性を持つ用言の格フレームから用例の集まりにくい用言の格フレームを補うことができる．図 4.4 に実際に補った用例から構築された格フレームの例を示す．

弾く/ひく:動 1

＜ガ格＞	人/ひと:32, 私/わたし:29, 誰/だれ:28, 彼/かれ:23, 先生/せんせい:17, 僕/ぼく:15, ...
＜ヲ格*＞	ギター/ぎたー:5519, ピアノ/ぴあの:5010, ベース/べーす:1055, 楽器/がっき:721, ...
＜ニ格＞	実際/じっさい:20, 一心不乱/いっしんふらん:8, 為/ため:7, よう/よう, ...
＜デ格＞	バンド/ばんど:88, 前/まえ:51, バック/ばっく:41, 中/なか:26, 趣味/しゅみ:22, ...

弾く/ひく:動 2

＜ガ格＞	人/ひと:6, 私/わたし:4, 誰/だれ:3, 三味線/しゃみせん:2, 先生/せんせい:2, ...
＜ヲ格*＞	曲/きょく:1575, ソロ/そろ:466, メロディー/めろでぃー:160, 旋律/せんりつ:80, ...
＜ニ格＞	実際/じっさい:7, よう/よう:3, 順番/じゅんばん:3, アンコール/あんこーる:2, ...
＜デ格*＞	ピアノ/ぴあの:245, 右手/みぎて:10, 前/まえ:6, 指/ゆび:5, ギター/ぎたー:5, ...

...

くれる/くれる:動 1

＜ガ格＞	人/ひと:199, 友人/ゆうじん:86, 友達/ともだち:76, 彼/かれ:49, 先生/せんせい:36, ...
＜ヲ格*＞	メール/めーる:2952, 電話/でんわ:1611, 返事/へんじ:1216, 連絡/れんらく:1207, ...
＜ニ格*＞	私/わたし:2112, 僕/ぼく:839, 俺/おれ:611, 御土産/おみやげ:206, 皆/みんな:85, ...
＜デ格*＞	無料/むりょう:132, メール/めーる:76, おまけ/おまけ:56, サービス/さーびす:32, ...

くれる/くれる:動 2

＜ガ格＞	友人/ゆうじん:3, あなた/あなた:2, 先生/せんせい:2, 人/ひと:2, キャラクター/きゃらくたー:1, ...
＜ヲ格*＞	ヒント/ひんと:498, 示唆/しさ:9
＜ニ格＞	私/わたし:13, あなた/あなた:2, 待ち/まち v:2, ナゾ/ナゾ:1, 我々/われわれ:1, ...

...

図 4.4: 補完された用言の格フレーム

第5章　格フレームを用いたかな表記語の曖昧性解消

かな表記語はJUMANによって日常使用の範囲で全ての可能性のある代表表記を出力される[1]．格解析によって選択される格フレームを参照し，それらの中から最も適当なものを選ぶことで，かな表記語の曖昧性解消を実現した．

かな表記語の曖昧性には，曖昧性のある語が用言である用言曖昧性と，曖昧性のある語が名詞である名詞曖昧性と二種類ある．以下では用言の格フレームの選択方法について述べた後，それぞれの曖昧性について曖昧性解消の手順を述べる．いずれも用言と格構造を入力とする．

さらに，代表表記化された格フレーム辞書と既存の格フレーム辞書の両方で曖昧性解消実験を行い，代表表記化による解析精度の向上について検討した．

5.1　曖昧性解消手法

5.1.1　格フレームの選択

KNPの格解析において，用言の格フレームが選択される．用言の用法の決定に対して，直前格要素が重要であり，特にヲ格，ニ格の場合にその傾向が強い．また，直前格要素が＜主体＞の場合，用言の用法は決まりにくい．これらを考慮して，以下の条件で格フレームの選択を行う．

1. 入力側の対象用言が直前格 *cm* と直前格要素 *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．

[1]例えば,「こしょう」には「呼称」,「故障」,「湖沼」,「胡椒」,「古称」,「孤称」,「誇称」,「扈従」,「股掌」,「胡床」,「胡牀」,「小姓」,「小性」と非常に多くの漢字表記があるが，JUMANでは日常使用される「呼称/こしょう」,「故障/こしょう」,「湖沼/こしょう」,「胡椒/こしょう」のみを出力する．また,「コショウ」が入力である時は,「胡椒/こしょう」のみを出力する．

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，C と格フレームの cm の各用例との類似度のうちもっとも高いものとする．C と用例との類似度は代表表記化した分類語彙表を利用し，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．

5.1.2 用言曖昧性解消

用言の曖昧性は「はやい（早い，速い）」のような原形の読みが同じものだけでなく,「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．用言曖昧性解消は，全ての候補用言の中から格フレームを選択し，選択された格フレームの代表表記（または漢字表記）を参照することで行う．用言曖昧性解消の手順について例とともに示す．

(1) 皮をむく

という文を解析すると,「むく」には「向く/むく」と「剥く/むく」の二種類の代表表記の可能性が出力され，曖昧性がある．この両方の用言の格フレームの中から，入力の「むく」の格フレームを選択することで，選択された用言の代表表記に，入力の用言の代表表記を決定し，曖昧性を解消する．

「むく」には表 5.1 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/かわ」との類似度を計算する.「剥く/むく：1」のヲ格の「皮/かわ」とマッチし，類似度が最大となるため，格フレームは「剥く/むく：1」に決定される．こうして,「むく」の代表表記を「剥く/むく」に決定する．

5.1.3 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の格要素を参照し，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．係り先の用言の格フレームにおいて，曖昧性のある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/そば」と「蕎麦/そば」の二種類の代表表記が考えられ，曖昧性が存在する．

名詞曖昧性解消は以下の手順で行われる．

表 5.1:「むく」の格フレーム

用言	格	用例
向く/むく：1	ガ	気/き，目/め，足/あし，…
	ヲ	下/した，上/うえ，ソッポ/そっぽ，…
	ニ	人/ひと，方向/ほうこう，自分/じぶん，…
向く/むく：2	ガ	＜主体＞
	ニ	＜補文＞
…		
剥く/むく：1	ガ	女性/じょせい，武/ぶ，…
	ヲ	皮/かわ，林檎/りんご，梨/なし，…
	デ	技/わざ，ナイフ/ないふ，手/て，…
	ニ	＜補文＞
	ノ	じゃが芋/じゃがいも，林檎/りんご，…
剥く/むく：2	ガ	魔物/まもの，システム/しすてむ，狼/おおかみ，…
	ヲ	牙/きば
	ニ	人間/にんげん，飼い主/かいぬし，私/わたし，…
…		

表 5.2: 選択される「食べる」の格フレーム

用言	格	用例
食べる/たべる	ガ	私/わたし，人/ひと，誰/だれ，…
	ヲ	御飯/ごはん，ラーメン/らーめん，…，蕎麦/そば，…
	デ	屋/や，店/みせ，家/いえ，…

1. 曖昧性のある単語がそれぞれの候補の意味コードを全て持つと仮定して，用言との各格フレームとの類似度を取り，用言の格フレームを選択する．

2. 各候補代表表記と選択された格フレームのすべての格要素との類似度を計算し，各候補代表表記ごとに類似度の最大値を求める．その最大値が最大となる代表表記に名詞の代表表記を決定し，曖昧性を解消する．

例を用いて説明すると,「そば」が「傍/そば」と「蕎麦/そば」の両方の意味コードを持つとして,「食べる/たべる」の各格フレームと類似度を計算し,「食べる」の格フレームが選択される．選択された格フレームのヲ格を表 5.2 に示す．

次に，曖昧性のある名詞の各候補である「蕎麦/そば」と「傍/そば」のそれぞれと,「食べる/たべる」のヲ格の格要素すべてについて類似度を計算する．すると，選択された格フレームの格要素に「蕎麦/そば」があるため，候補「蕎麦/そば」の

表 5.3: 曖昧性解消実験の結果

		既存格フレーム	代表表記化格フレーム
Web テキスト	用言曖昧性解消	75.0% (30/40)	84.8% (78/92)
675 文	名詞曖昧性解消	100.0% (4/4)	100.0% (3/3)
料理テキスト	用言曖昧性解消	76.4% (252/330)	83.6% (551/659)
7,800 文	名詞曖昧性解消	76.7% (33/43)	77.8% (35/45)

類似度は 1 となる.「傍/そば」は類似度が 1 となる格要素がないため,「蕎麦/そば」の類似度が最大となり，代表表記が「蕎麦/そば」に決定する．

5.2 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．格フレーム辞書は本研究で構築した代表表記化格フレーム辞書と既存の格フレーム辞書との 2 種類で実験を行った．使用したテキストは Web テキスト 675 文と，料理テキスト 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，格解析を行い，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．精度については，次のような尺度を用いている．

$$\text{精度} = \frac{\text{正しく曖昧性を解消できた単語}}{\text{解析によって曖昧性が解消された単語}}$$

曖昧性解消実験の結果を表 5.3 に示す．どの曖昧性解消に対しても既存格フレーム辞書を用いた時と比べて，代表表記化格フレーム辞書を用いた時の方が解析精度に向上が見られた．名詞曖昧性解消ではサンプルが少なく，大きな精度向上は見られないが，用言曖昧性解消では 7〜10%程度の精度向上が見られた．

また，代表表記化格フレーム辞書を用いることによって，大幅に曖昧性解消をされる用言が増えた．これは「成る/なる」など，日常でかな表記が一般的に用いられる用言の格フレームが，既存格フレーム辞書では十分に集まっておらず曖昧性解消ができていなかったためである．

以上の結果から，格フレーム辞書を代表表記化することの有効性を示すことができた．

代表表記化された格フレーム辞書を用いることで，以下のような解析誤りを防ぐことができた．○が代表表記格フレームを用いた解析で選ばれた正しい代表表

記であり，×が既存の格フレーム辞書を用いた誤った解析結果である．また，無印はその他の候補を表す．

1. ○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。

2. ○癌/がん　×ガン/がん　雁/がん
   【ガン】の予防に効果が期待できる。

3. ○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。

1 の例では，既存の格フレーム辞書では用例が十分でなかった「有る」の格フレームが,「有る」と「ある」をまとめて「有る/ある」の格フレームとなったことで用例が十分に集まり，正しく解析が行われるようになったと考えられる．

また，2 の例では「予防/よぼう」の格フレームの格要素の「ガン」の曖昧性が解消され,「癌/がん」と代表表記を決定しているために正しく解析が行われるようになったと考えられる．

さらに，3 の例では「弾く」という同一の用言の格フレームとして集められていた「弾く/ひく」と「弾く/はじく」を別々に扱えるようになったため，曖昧性の解消が行えるようになった．

第6章　まとめ

6.1　結論

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い，それらの持つ表記揺れや曖昧性の問題の解消を行った．また，それらを用いた格フレームによる曖昧性解消の実験を行い，本研究で構築された代表表記化格フレーム辞書を用いることによって，解析の精度が向上することを示した．

6.2　今後の課題

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が，人手で曖昧性を解消したり，補う用言を決定するといった作業が必要であり，再構築にかかるコストが大きいという問題があるため，このコストの低減について検討が必要であろう．

謝辞

本研究を進めるにあたって，二年間に渡り度重なるご指導と助言を頂いた江崎浩教授，黒橋禎夫教授に深く感謝致します．また，研究を進める上で技術的な指導をして頂き，支えて頂いた河原大輔氏に御礼申し上げます．そして，日々の研究活動の上で様々な手助けをして頂き，また，お互いに意見を交し高めあった研究室の先輩方と同期の皆様に感謝致します．

皆様，本当にありがとうございました．

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] 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)

[4] 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)

[5] Jack Halpern, “Lexicon-based Orthographic Disambiuation in CJK Intelligent Information Retrieval,” COLING 2002, (2002)

[6] 黒橋禎夫, 河原大輔, “日本語形態素解析システム JUMAN version 5.1 使用説明書”, 東京大学大学院情報理工学系研究室, (2005)

[7] 黒橋禎夫, 河原大輔, “日本語構文解析システム KNP version 2.0 使用説明書”, 東京大学大学院情報理工学系研究室, (2005)

[8] NTT コミュニケーション科学研究所, “日本語語彙大系”, 岩波書店, (1997)

[9] 国立国語研究所, “分類語彙表 増補改訂版” 大日本図書, (2004)

[10] L.R.Chapman, “Roget’s International Thesaurus (Fourth Edition),” Harper & Row, (1984)

[11] G. A. Miller, R. Bechwith, C. Fellbaum, D. Gross, K. Miller, and R. tengi, “Five papers on WordNet,” Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton University, (1993)

[12] 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)

[13] Masaki Murata et al., ”Japanese word sense disambiguation using the simple Bayes and support vector machine methods,” Proc. SENSEVAL-2, (2001)

[14] Donald Hindle, “NOUN CLASSIFICATION FROM PREDICATE-ARGUMENT STRUCTURES,” 28th Annual Meeting of the Association for Computational Linguistics, pp.268-275 (1990)

[15] Dekang Lin, “Automatic Retrieval and Clustering of Similar Words,” COLING-ACL, pp.768-774 (1998)

[16] Kentaro Torisawa, “An Unsupervised Learing Method for Associative Relationships between Verb Phrases” COLING’02, pp.1009-1015 (2002)

[17] Takenobu Tokunaga, Makoto Iwayama, Hozumi Tanaka, “Automatic thesaurus construction based on grammatical relations,” Proc. of IJCAI-95, pp.1308-1313 (1995)

[18] David Yarowski, “UNSUPERVISED WORD SENSE DISAMBIGUATION RIVALING SUPERVISED METHODS,” In Proceedings of the Annual Meeting of the Association for computational Linuistics, pp.186-196 (1995)

発表文献

[1] 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)

[2] 岡部浩司, 河原大輔, 黒橋禎夫, “代表表記による自然言語リソースの整備”, 言語処理学会第 13 回年次大会, (2007)