

格パターンの多様性に頑健な 日本語格フレーム構築

2015.12.04 @IPSJ-SIGNL224(名工大)

林部祐太(京大/CREST), 河原大輔(京大),
黒橋禎夫(京大/CREST)

格フレーム

述語ごとに以下が同じ用例をまとめたもの

1. 格の取り方 (格パターン)

2. 述語の語義

	格	項		格	項
頼る1	ガ	母, 親...	積む1	ガ	選手, 私, 学生, ...
	ヲ	子育て, 育児, ...		ヲ	経験, 実績, ...
	ニ	実家, 祖母, ...		デ	業界, 会社, ...
頼る2	ガ	母, 父, ...	積む2	ガ	ドライバー, 男, ...
	ヲ	親, 祖母,		ヲ	荷物, コンテナ,
	デ	子育て, 育児, ...		ニ	車, バイク, ...

母は実家へ電話した。
子育てを頼るようだ。

頼る2

母が子育てを実家に

構文解析, 述語項構造解析に有用

提案手法の概要

従来の自動格フレーム構築手法:

複数の格パターンをもつ述語をうまく扱えないことがある

	格	項
拡大する1	ガ	エリア, 区域, ...
	ヲ	エリア, 地域, 区域, ...
	ニ	全国, 全域, ...

X

提案手法: 格パターンの多様性に頑健な自動構築手法

	格	項
拡大する1	ガ	エリア, 区域, ...
	ニ	全国, 全域, ...
拡大する2	ガ	ドコモ, AU
	ヲ	エリア, 地域, 区域, ...
	ニ	全国, 全域, ...

✓

従来の自動格フレーム構築手法

従来手法1: [河原+02,05]

1. 大規模テキストを形態素・構文解析
2. 解析誤りの可能性が低い述語項構造を抽出
3. **直前項**が共通しているもの同士でまとめる(初期格フレーム)
4. 初期格フレームを(一度に)クラスタリング

従業員	が	車	に	荷物	を	積む	初期格フレーム1
太郎	が			荷物	を	積む	

彼	が	車	に	物資	を	積む	初期格フレーム2
---	---	---	---	----	---	----	----------

従業員	が			経験	を	積む	初期格フレーム3
選手	が			経験	を	積む	

従来手法2:[河原+02,05]の改善

- [河原+02,05]の問題点

- 項の出現順が異なれば別初期格フレームとなり，同一格フレームに属さない可能性がある

荷物を 車に 積む

車に 荷物を 積む

- 改善点

- 全用例から述語の直前項の割合を予め調べる
- その割合に基づく順序に従って，初期格フレームを決める格を選ぶ

荷物を 車に 積む

車に 荷物を 積む

「積む」の直前格ランキング

1.ヲ格 2.二格 3.へ格 4.マデ格

従来手法の問題点

- 格パターンが異なる初期格フレームが，クラスタリングの結果，同じクラスに属することを完全に防げない

初期格フレーム1

エリアが 全国に 拡大する
 エリアが 全域に 拡大する
 エリアが 地方に 拡大する
 ...

初期格フレーム2

ドコモが エリアを 全国に 拡大する
 AUが エリアを 全域に 拡大する
 ドコモが エリアを 地方に 拡大する
 ...



	格	項
拡大する1	ガ格	エリア, 区域, ...
	ヨ格	エリア, 地域, 区域, ...
	二格	全国, 全域, ...



複数の格パターンをもつ述語の例

複数の格パターンをもつ述語の例(1/2)

- 動詞

- 拡大する（自他同形動詞）

- エリア{が,を}拡大する

- 飲める（可能動詞）

- 今日は 酒{が,を} 飲める

- 走る

- 地震で 亀裂が 地面{に,を} 走った

- 感染する

- 学生が インフルエンザに 感染する

- 学生に インフルエンザが 感染する

複数の格パターンをもつ述語の例(2/2)

- “動詞+接尾辞”
 - 接尾辞「やすい」
 - パン{が,を} 食べやすい
 - 接尾辞「いる」
 - インターネットで ソフトウェア{が,を} 売っている
- “テ形+接尾辞”(い抜き言葉)
 - インターネットで ソフトウェア{が,を} 売ってます

接尾辞が付属することにより格パターンが増えることがある
→原則、後続する接尾辞列が異なれば別の述語として扱う

格フレーム述語単位の定義

- 原則，後続接尾辞列が異なれば別の述語として扱う
- ただし，テ形に続く場合を除き，「**ます**」の有無は区別しない

述語代表表記	例
飾る	飾る，飾り ます
飾る+れる	飾られる，飾られ ます
飾る+れる+いる	飾られている，飾られて います
飾る+いる	飾っている，飾って います
飾る(テ形)+ます	飾ってます

クラスタリングに用いる用例数が減る
→十分な量のテキスト(40億文超)があるので問題無

提案手法

提案手法

用例を3種類に分け,それぞれ格フレームを構築し,最後にそれらを統合する

(1) ヲ格を含む用例

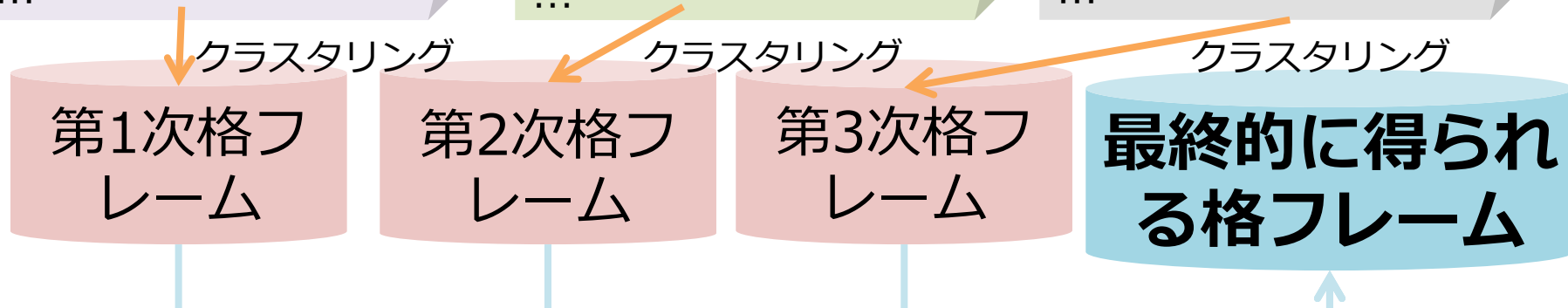
- 規模を拡大する
- 規模を倍に拡大する
- エリアを拡大する
- ドコモがエリアを拡大する
- ...

(2) ヲ格は含まず無生物であるガ格を含む用例

- 規模が拡大する
- 規模が急に拡大する
- エリアが拡大する
- エリアが倍に拡大する
- ...

(3) その他の用例

- <数量>%に拡大する
- 前年比で<数量>%に拡大する
- 会員が<数量>人に拡大する
- ...



格	項
ガ ヲ	ドコモ, AU エリア, 地域, ...

格フレームの例

格	項
ガ	エリア, 地域, ...

格フレームの例


	格	項
拡大する1	ガ格 二格	エリア, 区域, ... 全国, 全域, ...
拡大する2	ガ格 ヲ格	ドコモ, AU エリア, 地域, 区域, ...

統合時の処理

1. ガ格の省略されている格フレームの削除

格	項
ガ	イベント, ...
ニ	盛況, ...

格	項
ニ	盛況, ...



2. 格交替の認識

格	項
ガ	ドコモ, AU
ヲ	エリア, 地域, ...
ニ	全国, 全域, ...

格	項
ガ	エリア, 地域, ...
ニ	全国, 全域, ...



(再掲)提案手法

(1)ヲ格を含む用例

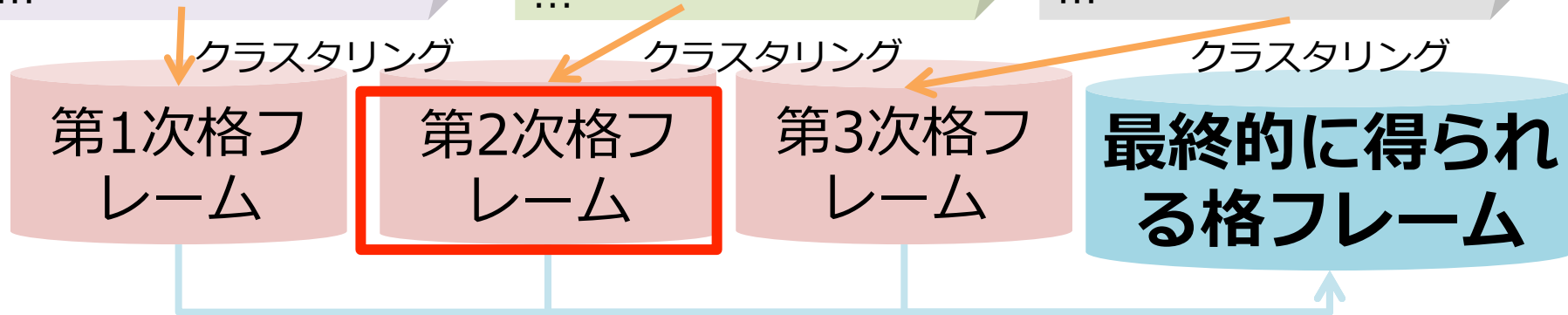
- 規模を拡大する
- 規模を倍に拡大する
- エリアを拡大する
- ドコモがエリアを拡大する
- ...

(2)ヲ格は含まず無生物であるガ格を含む用例

- 規模が拡大する
- 規模が急に拡大する
- エリアが拡大する
- エリアが倍に拡大する
- ...

(3)その他の用例

- <数量>%に拡大する
- 前年比で<数量>%に拡大する
- 会員が<数量>人に拡大する
- ...



第2次格フレームの種類

(1) 第1次格フレーム(他動詞用法格フレーム)と格交替する格フレーム

例: {シェア, ...}が 拡大する

1次格フレーム: {企業, ...}が {シェア, ...}を 拡大する

例: {痛み, ...}が {背中, ...}に 走る

1次格フレーム: {痛み, ...}が {背中, ...}を 走る

(2) 第1次格フレームのヲ格を省略した用例からなる格フレーム

A: ヲ格の省略が不自然な場合

例: {遠赤外線, ...}が 温める

1次格フレーム: {遠赤外線, ...}が {体, ...}を 温める

削除したいがBとの区別が困難

B: ヲ格の省略が自然な場合

例: {花, ...}が 開く

1次格フレーム: {花, ...}が {つぼみ, ...}を 開く

(3) その他

(再掲)提案手法

(1)ヲ格を含む用例

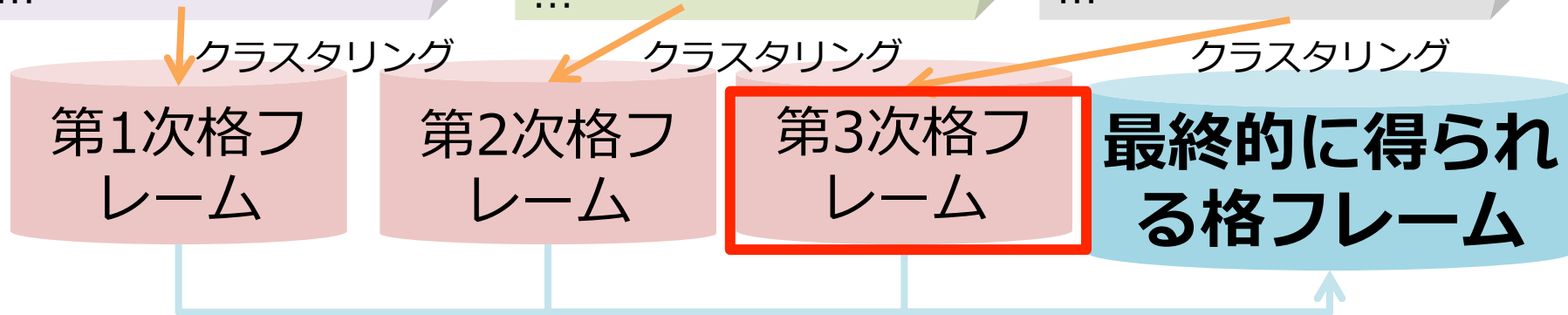
- 規模を拡大する
- 規模を倍に拡大する
- エリアを拡大する
- ドコモがエリアを拡大する
- ...

(2)ヲ格は含まず無生物であるガ格を含む用例

- 規模が拡大する
- 規模が急に拡大する
- エリアが拡大する
- エリアが倍に拡大する
- ...

(3)その他の用例

- <数量>%に拡大する
- 前年比で<数量>%に拡大する
- 会員が<数量>人に拡大する
- ...



第3次格フレームの種類

(1) 第1次格フレーム(他動詞用法格フレーム)と格交替する格フレーム

例: {子供, ...} が {プール, ...} で 泳ぐ

1次格フレーム: {子供, ...} が {プール, ...} を泳ぐ

(2) 第1次格フレームのヲ格を省略した用例からなる格フレーム

A: ヲ格の省略が不自然な場合

削除したいがBとの区別が困難

例: {彼, ...} が {実行, ...} に 移す

1次格フレーム: {彼, ...} が {計画, ...} を {実行, ...} に移す

B: ヲ格の省略が自然な場合

例: {彼, ...} が {渋谷, ...} で 飲む

1次格フレーム: {彼, ...} が {渋谷, ...} で {酒, ...} を飲む

(3) 第2次格フレームのガ格を省略した用例からなる格フレーム

例: {盛況, ...} に 終了する

2次格フレーム: {イベント, ...} が {盛況を ...} に終了する

削除

(4) その他

(再掲)提案手法

(1) フラグを含む用例

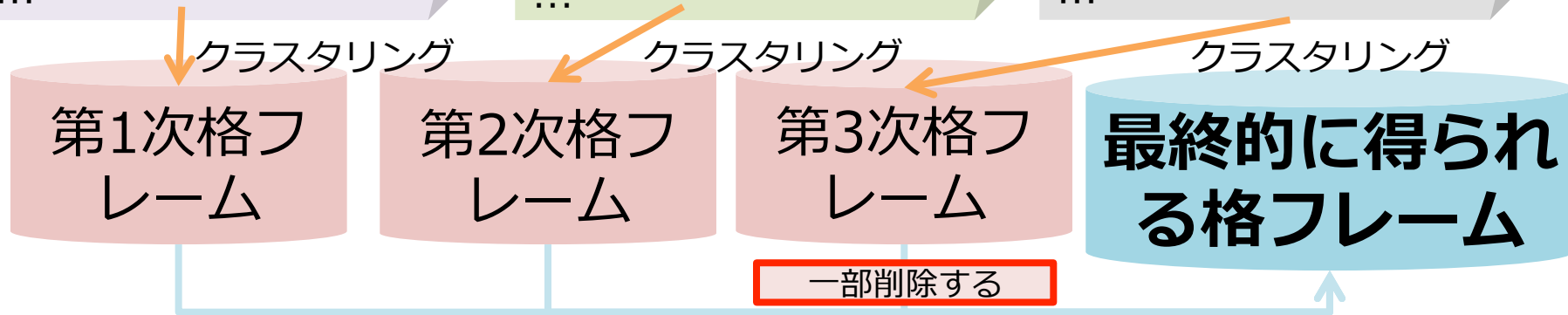
- 規模を拡大する
- 規模を倍に拡大する
- エリアを拡大する
- ドコモがエリアを拡大する
- ...

(2) フラグは含まず無生物であるガ格を含む用例

- 規模が拡大する
- 規模が急に拡大する
- エリアが拡大する
- エリアが倍に拡大する
- ...

(3) その他の用例

- <数量>%に拡大する
- 前年比で<数量>%に拡大する
- 会員が<数量>人に拡大する
- ...



評価実験

格フレーム構築の設定

- webから収集した日本語文をフィルタリングして得た約40億文
- JUMANとKNPで形態素・構文解析
- 約4万述語に対して述語項構造を抽出
- 平均格フレーム数6.9個
- クラスタリングには, Chinese Restaurant Process に基づく手法 [Kawahara+14]を一部改良した手法を用いた

提案手法の有無の比較実験

- 評価対象：20述語の各頻度上位10格フレーム
 - 自他同形動詞：拡大する,増加する,閉じる,決定する,完成する
 - 母音動詞+られる：食べられる,見られる,当てられる,設けられる,飛ばされる
 - 可能動詞：使える,折れる,消せる,行ける,打てる
 - 普通の動詞：積む,飾る,読む,味わう,抜く
- 評価方法：妥当な格フレーム数の人手比較

「妥当な格フレーム」の基準

- 各格に不自然な項が混ざっていない
- ガ格, ヲ格, 二格の項を組み合わせた用例が許容できる

	格	項
拡大する1	ガ	エリア, 区域, ...
	ヲ	エリア, 地域, 区域, ...
	二	全国, 全域, ...

X

実験結果: 妥当な格フレーム数の変化

述語	提案手法なし	提案手法あり	差分
見られる	3	10	+7
当てられる	3	10	+7
拡大する	3	9	+6
決定する	4	10	+6
使える	3	8	+5
増加する	4	7	+3
飛ばされる	7	10	+3
食べられる	8	10	+2
閉じる	9	10	+1
完成する	9	10	+1
設けられる	9	10	+1
折れる	9	10	+1
打てる	9	10	+1
飾る	9	10	+1
抜く	8	8	+0
(その他5述語)	10	10	+0
計	145/200	191/200	+46

平均格フレーム数:
39.5 → 46.3

提案手法による改善例(1/2)

✕	拡大する:9	ガ格 140 ヲ格 3060 ニ格 511	対象 130, 事業所 7, 〈数量〉区 3 対象 2077, 一部 405 対象者 223, … サイズ 143, {数量}pix 52, ウィンドー 33, 画
✓	拡大する:10	ガ格 3 ヲ格 2806 ニ格 353	〈数量〉区 3 対象 2077, 対象者 223, 適用 190, 助成 159, 輪 74, 募集枠 39, … 〈数量〉人 29, 〈数量〉年生 27, 〈数量〉回 14, 全域 13, 商品 12
✓	拡大する:18	ガ格 186 ニ格 10	対象 176, 助成 10 資産 4, 業種 3, 〈数量〉回 3

提案手法による改善例(2/2)

評価	格フレーム	格	用例
✕	当てられる:1	ガ格 215 ヲ格 96 ニ格 1120	手 73, 手の平 20, 刃 18, 指 18, 唇 15, … 手 35, 聴診器 12, ボール 9, ナイフ 9, ローター 6, … 毒気 187, 頬 149, 首筋 131, 顔 89, 背中 71, 胸 70, 唇 62, 首 57, …
✓	当てられる:2	ガ格 5 ニ格 664	一部 5 費用 224, 資金 142, 返済 78, 支払い 62, 経費 61, 運営費 53, 活動



評価	格フレーム	格	用例
✓	当てられる:3	ヲ格 158 ニ格 74	手 128, 指 18, 人差し指 12 額 19, 頬 12, 唇 9, 口 8, 胸 6
✓	当てられる:5	ガ格 225 ニ格 106	手の平 61, 指 57, 唇 42, 人差し指 30, ナイフ 24, 指先 唇 32, 額 22, 首筋 16, 頬 15, 上 5, 背中 4, …

改善できなかつた例 (1/2)

1. ヲ格を省略した格フレームを削除できていない → ヲ格の必須性の判断が必要

使える:2	ガ格 8	アリス 5, 自分 3
	ニ格 22562	時 13440, 実際 2178, 人 1438, 事 1022, ...

改善できなかった例 (2/2)

x	拡大する:1	ガ格 13	おら 9, 画像 4
		ヲ格 17081	写真 15725, 画像 944, イメージ 236, 映像 84, ...
		ニ格 38	〈数量〉倍 24, 大きさ 5, ...

2. 形態素解析の誤り→語彙の増強

- 「おらが村の健康茶減肥茶の写真を拡大する」から「おら」をガ格として解釈

3. 係り受け解析の誤り→確信度付与ルールを修正

- 「画像をクリックで画像が拡大します。」からガ格とヲ格ともに「画像」を抽出

まとめ：本論文の貢献

1. 格フレームの述語単位を定義した
2. 格パターンの多様性に頑健な日本語格フレーム構築手法を提案した
 - 主観評価で大幅に改善
 - 今後はこれを使って述語項構造解析
3. (本発表では割愛) 可能動詞をJUMAN辞書に網羅的に追加した

構築した格フレームは公開予定

APPENDIX

JUMAN辞書への可能動詞の追加

- JUMAN辞書では主要な可能動詞は登録されていた
- 多くの表現は適当な文脈を与えれば可能形を作ることができる
 - 「あの映画で涙ぐめる若さが羨ましい」
- 今回、原則的に全ての可能動詞を自動生成して機械的に登録するように変更

辞書追加しない可能動詞

可能形として用いられることが極めて考えにくいにも関わらず、形態素解析に曖昧性を増やしてしまう場合追加しない

1. 「別の子音動詞+接尾辞れる」

- 例1: 「刺さる」の可能動詞「刺される」
（「刺す+れる」と解釈できる）
- 例2: 「抜かる」の可能動詞「抜かれる」
（「抜く+れる」と解釈できる）

2. 「別の子音動詞+接尾辞せる」

- 例1: 「切らす」の可能動詞「切らせる」
（「切る+せる」と解釈できる）

3. 別の一語の「動詞」

- 例1: 「泡立つ」の可能動詞「泡立てる」
- 例2: 「間違う」の可能動詞「間違える」