Vol.2015-NL-224 No.14
2015/12/4

# 格パターンの多様性に頑健な日本語格フレーム構築

林部 祐太[1,2,a)]　河原 大輔[1,b)]　黒橋 禎夫[1,2,c)]

**概要**：述語項構造のパターンである格フレームは構文・格解析，さらに照応・省略解析における重要な知識源である．従来の格フレーム構築の研究では，用例をクラスタリングして，述語の語義と格パターンの違いの両方をとらえようとしてきたが，格パターンの多様性に頑健ではないという問題があった．本稿では，用例を格の出現の仕方に基づいて 3 種類に分け，それぞれについて格フレームを構築してから最後に統合する手法を提案する．また，格パターンを正しく認識するために必要な格フレームの述語単位の定義と，可能動詞・サ変動詞の語彙の整備も行う．評価実験では，日本語ウェブテキスト 40 億文に対して提案手法を用いた格フレームを構築し，従来手法と比較して妥当な格フレームの割合が 72.5%(145/200) から 95.5%(191/200) に大きく改善したことを確認した．

## 1. はじめに

自然言語の理解のためには，単語間の意味的関係を抽出することが重要である．特に，述語とその項の関係は，最も基礎的な関係の 1 つであり，その解析は述語項構造解析，意味役割付与などと呼ばれている．項はしばしば省略されるが，その省略される項を補う省略解析も述語項構造解析の一部とみなせる．日本語においてこの解析は現状で最も困難だが重要な基礎解析であると考えられている．

述語項構造解析において重要な知識資源の 1 つに格フレームがある．格フレームとは，述語ごとに，述語の語義，および格のとり方が同じ用例をまとめたものである．

例えば「拡大する」は，自動詞用法と他動詞用法の 2 種類の格のとり方がある．他動詞用法は例 (1a) のようにガ格に対象物をとりヲ格はとらないパターンで，他動詞用法は例 (1b) のようにガ格に動作主体をとりヲ格に対象物をとるパターンである．

(1)　a. 規模が 倍に 拡大する
　　b. 企業が 規模を 倍に 拡大する

このような動詞は自他同形動詞とよばれ，漢語動詞に多く見られる．本稿では格のとり方を格パターンとよぶ．

従来の研究では，用例をクラスタリングして，述語の語義と格パターンの違いの両方をとらえようとしてきたが，異なる格パターンを含む格フレームを作ってしまうという問題があった．例えば，河原ら [1] は「述語と述語の直前項の組を単位にして考えると用法はほとんど一意に決定される」と仮定して，述語の直前項が共通しているもの同士で用例をまとめ，それを最小単位としてクラスタリングを行った．しかしながら，例 (1a) と例 (1b) の 2 つの用例を述語の直前にあるニ格の項でまとめてしまうと，

(2)　{ 規模, . . .} が { 規模, . . .} を { 倍, . . .} に 拡大する

という誤った格フレームができてしまう．日本語では語順が自由であるため，単純に述語の直前項を用いてクラスタリングの単位としてしまうと，実際には用いられない格の組み合わせをもつ格フレームができる．

ガ格とヲ格が交替した両方のパターンが存在することによる同様の問題は，受身形，可能形，接尾辞「やすい」，「いる」が接続する場合等でも起こる．

(3)　受身形
　　a. 企業で 技術が 使われる
　　b. 企業で 技術を 使われる

(4)　可能形
　　a. 今日は 酒が 飲める
　　b. 今日は 酒を 飲める

(5)　接尾辞「やすい」
　　a. パンが 食べやすい
　　b. パンを 食べやすい

(6)　接尾辞「いる」
　　a. インターネットで ソフトウェアが 売っている

[1] 京都大学 大学院情報学研究科
[2] 科学技術振興機構 CREST
[a)] yuta-h@i.kyoto-u.ac.jp
[b)] dk@i.kyoto-u.ac.jp
[c)] kuro@i.kyoto-u.ac.jp

b. インターネットで ソフトウェアを 売っている

また，以下のように，ガ格とヲ格の他でも格交替による同様の問題が存在する．

(7) ニ格とヲ格の交替
a. 地震で 亀裂が 地面に 走った
b. 地震で 亀裂が 地面を 走った

(8) デ格とヲ格の交替
a. 魚が 清流で 泳ぐ
b. 魚が 清流を 泳ぐ

(9) ニ格とガ格の交替
a. 学生が インフルエンザに 感染する
b. 学生に インフルエンザが 感染する

そこで，用例を格の出現の仕方に基づいて分け，それぞれについて格フレームを構築し，最後にそれらを統合する格パターンの多様性に頑健な手法を提案する．

## 2. 関連研究

### 2.1 日本語格フレーム

著者らは大規模なウェブコーパスを自動解析した結果から日本語格フレームを構築する手法を提案した [1], [2]．その手法では，述語と述語の直前項の組をもとに最初のクラスタを作り，それらを段階的にマージしていくことで格フレームを構築した．そして，得られた格フレームを用いて，構文・格解析の精度を改善した [3]．また，格フレームの規模を大きくするほど述語項構造解析に効果が出ることが [4] によって示されている．

### 2.2 英語格フレーム

人手で構築した英語の格フレームには，FrameNet [5] や PropBank [6] があり，それらは色々な研究で使われている．

テキストから自動で構築した研究には LDA-Frames[7], [8] や，Chinese Restaurant Process [9] を用いる方法 [10] がある．

LDA-Frames は．British National Corpus (BNC) から (subject, verb, object) の 3 つ組を抽出し，Latent Dirichlet Allocation (LDA) と Dirichlet Process を使うことで構築される．英語では項の省略が起こらないため，この手法を日本語には直接適用できない．

### 2.3 Chinese Restaurant Process に基づく格フレーム構築 [10]

本稿の評価実験ではベースラインとして [10] を用いるため，詳述する．

[10] はテキストから抽出した項構造を，“predominant argument”[*1] でまとめ，それをクラスタリングの最小単位とした．本稿ではその最小単位を初期格フレームとよぶ．

そして，初期格フレーム $v_i$ が格フレーム $f_j$ に属する事後確率 $P(f_j \mid v_i)$ を式 1 のように Chinese Restaurant Process で定義し，ギブスサンプリングを行うことで，各初期格フレームが最終的に所属する格フレームを求めた．

$$P(f_j \mid v_i) \propto \frac{n(f_j)}{N+\alpha} \cdot P(v_i \mid f_j) \tag{1}$$

第 1 項はディリクレ過程の事前分布で，第 2 項は $v_i$ の尤度である．$N$ は初期格フレームの数を示し，$n(f_j)$ は現在 $f_j$ に属する初期格フレームの数を示す．もし $f_j$ が新しい格フレームならば，$n(f_j) = \alpha$ とする．$\alpha$ は最終的に出来る格フレームの数に影響する．

尤度 $P(v_i \mid f_j)$ はディリクレ多項分布で定義される．

$$P(v_i \mid f_j) = \prod_{w \in V} P(w \mid f_j)^{count(v_i, w)} \tag{2}$$

ここで，$\mid V \mid$ は全格フレーム中の語彙の異なり総数である．ただし，“subj:bread” と “obj:bread” 等，格が異なれば別の語として数える．

$P(w \mid f_j)$ は次式のように定義した．

$$P(w \mid f_j) = \frac{count(f_j, w) + \beta}{\sum_{t \in V} count(f_j, t) + \mid V \mid \cdot \beta} \tag{3}$$

$count(f_j, w)$ は格を区別した単語 $w$ の $f_j$ における頻度，$\beta$ はディリクレ分布のハイパーパラメータである．

## 3. 格フレームの述語単位の定義と動詞の語彙整備

### 3.1 格フレームの述語単位の定義

著者らはこれまで，述語が受身形，使役形，「〜もらう」，「〜たい」，「〜ほしい」，「〜できる」の形であれば，格交替が起こり格と項の関係が通常の場合と異なるとして，能動形とは区別して異なる述語として扱ってきた [1]．しかしながら，その他にも格交替が起きる表現がある．

そこで，格フレームの述語単位を再定義する．その定義に基づく述語の例を表 1 に示す．本稿では形態素解析器に JUMAN を用いるため，以下では JUMAN が採用している益岡・田窪文法 [11] に基づく品詞体系の用語を用いて述べる．

**後続する接尾辞の区別**

例えば接尾辞「やすい」「いる」は例 (5) や例 (6) のように，ガ格とヲ格の交替を起こすことがある．このような格交替を起こす接尾辞を網羅することは困難であるため，本稿では動詞に後続する接尾辞の列が異なれば，原則として全て異なる述語として扱う．例えば，「認めたい」，「認めら

[*1] “dobj”, “ccomp”, “nsubj”, “prep ∗”, “iobj” の順序のうち，その述語項構造において最も高い順位をもつ項

| 述語 | 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，飾ってはいる，飾ってはいます，飾ってもいる，飾ってはいます |
| 飾る (テ形)+ます | 飾ってます |
| 飾る+こと+できる | 飾ることができる，飾ることはできる，飾ることもできる |

**表 1** 述語の例

| 種類 | 活用型 | 可能形 | 例 |
|---|---|---|---|
| 1 | 子音動詞 | 語幹+eru | 読める，切れる |
| | 母音動詞 | 未然形+接尾辞「られる」 | 寝られる，着られる |
| | サ変動詞 | 語幹+できる/せる [*2] | 左右できる，愛せる |
| | カ変動詞 | 語幹+接尾辞「られる」 | 来られる |
| 2 | 全動詞共通 | 基本形 + こと + { が，は，も } + { できる，出来る，可能だ，不可能だ } | 読むことができる，寝ることも可能だ |

**表 2** 2 種類の動詞の可能形の作り方

れる」,「認められたい」は全て異なる述語として扱う．なお,「飾ってはいる」のように間に助詞を挿入することができるが，そのような助詞は無視して考える．例えば,「飾っている」と「飾ってはいる」は同じ述語として考える．

例外とする接尾辞は「ます」である．「飾ってます」のように，動詞のテ形に「ます」が続く場合 [*3] を除いて,「ます」の有無は格交替に影響しないことが明らかなので，そのような「ます」は無視して考える．例えば,「飾る」と「飾ります」,「飾っている」と「飾っています」はそれぞれ同じ述語として扱うが,「飾る」と「飾ってます」は異なる述語として扱う．

このように後続する接尾辞を全て区別することで，格交替が起きる述語と起こらない述語が混ざる危険が無くなる．クラスタリングに用いる用例数が減る問題はあるが，十分な量のテキストがあれば影響はないと考えられる．

**可能形を作る表現の区別**

可能形は例 (4) のように格交替を起こすことがあるため，能動形との区別が必要である．可能形の作り方は，表 2 のように，その動詞の活用型によって異なる作り方と，全活用型共通の作り方の 2 種類がある．

動詞の活用型によって異なる作り方では，母音動詞とカ変動詞は語幹に接尾辞「られる」[*4] を接続する．そのため，前述した接尾辞の区別により，能動形と可能形は区別できる．一方，子音動詞とサ変動詞は動詞自体が別の母音動詞（可能動詞）に変わる．したがって，可能形を正しく認識するためには，可能動詞を形態素解析辞書に登録する必要がある．そこで，JUMAN 辞書の可能動詞の整備を行った．これについては，3.2 節で詳述する．

全活用型共通の作り方では，動詞の基本形に

こと + { が，は，も } + { できる，出来る，可能だ，不可能だ }

を続ける．そのため，この場合も能動形とは異なる述語として扱う．例えば,「食べる」と「食べることができる」と「食べることが不可能だ」は全て異なる述語として扱う．なお，助詞が「が」,「は」,「も」のいずれがあっても，同一の述語として扱う．例えば,「食べることができる」と「食べることはできる」は同じ述語として扱う．

### 3.2 動詞の語彙整備

**可能動詞の追加**

従来の JUMAN 辞書では，主要な可能動詞は登録されていたが，意志性を感じにくいもの等，可能形が作りにくいと考えられるものは登録されていなかった．

しかし，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．例えば,「涙ぐむ」は自然発生的に涙を出す行為を表し，通常は意志をもってすることができないが，

(10) あの映画で涙ぐめる若さが羨ましい

のように，文脈次第で可能形を使うことができる．

そこで，多様な表現を解析できるように，原則として全ての可能動詞を自動生成して機械的に登録することにした．ただし，実際には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以下のような場合は，登録から除外することにした．

- 「別の子音動詞+接尾辞れる」と解釈できるもので可能形と解釈しにくいもの
  例:「刺さる」の可能動詞「刺される」(「刺す+れる」と解釈できる),「抜かる」の可能動詞「抜かれる」

[*3] 話し言葉を中心に用いられる表現で，い抜き言葉とよばれる
[*3] 「愛する」のように，漢字一文字に「する」が付いた場合は「愛せる」のように「する」が「せる」になる． それ以外の場合は「結合する」が「結合できる」になるように,「する」が「できる」になる．
[*4] 接尾辞「られる」の代わりに接尾辞「れる」が話し言葉を中心に用いられることがある（ら抜き言葉).「られる」は可能の他に受身・自発・尊敬のヴォイスを付与するために使えるため，母音動詞とカ変動詞に接尾辞「られる」を続けたものには，ヴォイスの解釈に曖昧性がある．一方,「れる」は可能のヴォイスしか付与できないため，そのような曖昧性は生じない．

（「抜く+れる」と解釈できる）

- 「別の子音動詞+接尾辞せる」と解釈できるもので可能形と解釈しにくいもの
  例: 「切らす」の可能動詞「切らせる」（「切る+せる」と解釈できる）
- 別の一語の「動詞」の解釈があり可能形と解釈しにくいもの
  例: 「泡立つ」の可能動詞「泡立てる」,「間違う」の可能動詞「間違える」

**サ変動詞に対応するサ行子音動詞の追加**

「愛する」に対する「愛す」のように，サ変動詞に対応するサ行子音動詞は，よく使われるものは登録していた．しかし，動詞のカバレッジを上げるため，実際にはほぼ使われない動詞を除いてサ変動詞から元にサ行子音動詞とサ変動詞の可能形（可能動詞）を自動生成し，辞書に登録した．除外する動詞は以下の動詞である．

- サ行子音動詞も可能動詞も生成しない:
  する，幸い (さいわい) する，相対 (あいたい) する，相反 (あいはん) する
- サ行子音動詞を生成するが，可能動詞「〜せる」は自動生成しない:
  汗 (あせ) する，値 (あたい) する，心 (こころ) する，倍 (ばい) する，私 (わたくし) する
- サ行子音動詞を生成しないが，可能動詞「〜できる」を自動生成する:
  たむろする，どうかする，異 (こと) にする，左右 (さゆう) する，全う (まっとう) する，無 (む) にする

## 4. 格パターンの多様性に頑健な日本語格フレーム構築

### 4.1 3 種類の用例

クラスタリングの最小単位である初期格フレームを作る方法として，各用例の述語の直前項を利用する [1] の方法の他に，ある述語の全用例から述語の直前項の割合を予め調べておき，その割合に基づく順序に従って，用例が属する初期格フレームを決める格を選ぶ方法も考えられる．以下では，その格を弁別格，その順序を弁別格順序とよぶ．例えば「拡大する」の全用例における述語の直前項を頻度の降順で並べると

ヲ格 > ガ格 > デ格 > ...

という弁別格順序が得られ，各用例がもつ格のうち最も弁別格順序が高い格を弁別格とすると，例 (1a) はガ格，例 (1b) はヲ格が弁別格となり，別の初期格フレームに属すようになる．

しかし，自他同形動詞に対して，他の格がヲ格とガ格よりも高順位である弁別格順序を得てしまうと，得られる格フレームは格パターンを反映しない可能性が高い．また，例 (9a) と例 (9b) のような場合は，ニ格とヲ格のいずれが弁別格であっても，2 つの格パターンを区別した初期格フレームを作ることはできない．

そこで，格パターンが異なる初期格フレームが同じクラスタに属することを防ぐため，以下のように用例を次の 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手法を考案した（**図 1**）．

1. **第 1 次格フレーム（他動詞パターン格フレーム）**: ヲ格を含む用例を用いる．ヲ格の項がある用例は，他動詞パターンであることが明らかなので，ヲ格を弁別格として初期格フレームを作る．それらをクラスタリングすることで，他動詞パターンを示す格フレームを構築する．
2. **第 2 次格フレーム**: 次に，ヲ格を含まないが，非動作主体 [*5] のガ格の項を含む用例を用いる．例 (7b) の「亀裂」や例 (9b) の「インフルエンザ」は非動作主体のガ格の項である．これらの用例は自動詞用法の可能性が高いと考えられるのでガ格を弁別格として初期格フレームを作る．それらをクラスタリングすることで，自動詞パターンを示す格フレームを構築する．例 (1a) の「企業」，例 (8a) の「子供」，例 (9a) の「魚」のように動作主体がガ格項の場合，ガ格項は格パターンに必ずしも影響しないため，そのようなガ格項を含む用例はここでは用いない．
3. **第 3 次格フレーム**: ヲ格も，非動作主体のガ格の項も，どちらも含まない用例を用いる．ヲ格とガ格を除いた弁別格順序から弁別格を選び，初期格フレームを作り，クラスタリングし，格フレームを構築する．

### 4.2 3 つの格フレームの統合

ここで，3 種類の格フレームから最終的に作成する格フレームに採用すべき格フレームについて考える．全初期フレームの用例数の総和を $s_I$ とする．そして，3 種類の格フレームにある全ての格フレームを用例数の降順でソートして，先頭から順に頻度の和 $t$ を求めていき，$t \geq s_I \times 0.9$ となるまでの格フレームを採用検討対象とする．それ以外の格フレームはノイズを含む可能性が高いとして採用しない．また，採用検討対象としたものでも，別の格フレームのある格の省略と判断されるものは採用しない．

**第 1 次格フレーム（他動詞パターン格フレーム）**

第 1 次格フレームは，全て他動詞パターンであることが明らかである．ヲ格は格パターンの決定への影響力が強いため，採用検討対象となった格フレームは全て採用する．

**第 2 次格フレーム**

第 2 次格フレームは以下のように分類できる．

2-A ガ格をヲ格に交替させた第 1 次格フレームが存在する

[*5] JUMAN 辞書において，カテゴリが「人, 組織・団体, 動物」のいずれか，又は品詞細分類が「人名, 組織名」である単語を動作主体とし，それ以外の単語をさす

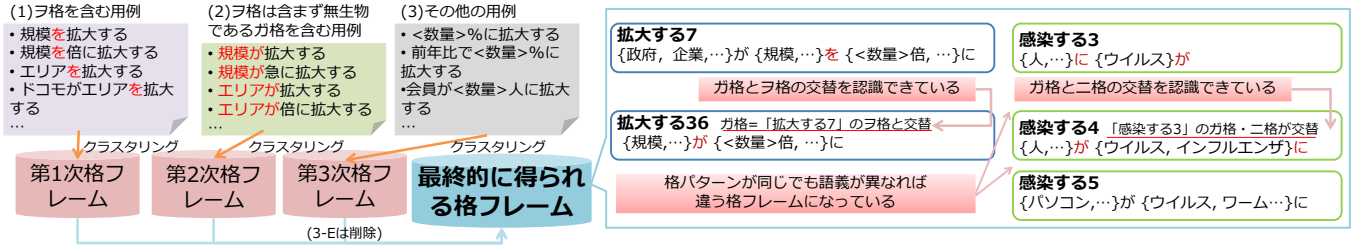


**図 1** 格パターンの多様性に頑健な日本語格フレーム構築

（狭義の自他同形の自動詞パターン）

(11) { シェア, . . . } が 拡大する
→ { 企業, . . . } が { シェア, . . . } を 拡大する

2-B ガ格**以外**（本稿ではデ格，ニ格を候補とする）をヲ格に交替させた第 1 次格フレームが存在する（広義の自他同形の自動詞パターン）

(12) { 痛み, . . . } が { 背中, . . . } に 走る
→ { 痛み, . . . } が { 背中, . . . } を 走る

2-C1 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と考えられる

(13) { 遠赤外線, . . . } が 温める
→ { 遠赤外線, . . . } が { 体, . . . } を 温める

2-C2 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とも，述語自体にヲ格相当のものが含意されている格パターンとも考えられる

(14) a. { 花, . . . } が 開く
→ { 花, . . . } が { つぼみ, . . . } を 開く
b. { 鯉幟, . . . } が 泳ぐ
→ { 鯉幟, . . . } が { 空, . . . } を 泳ぐ

2-D 2-A, 2-B, 2-C1, 2-C2 のいずれでもない（第 1 次格フレームと対応付かない自動詞パターン格フレーム）

(15) a. { 話, . . . } が 弾む [*6]
b. { ライト, . . . } が消える

これらのうち，採用したくないものは 2-C1 であるが，2-C1 と 2-C2 との区別を付けることは難しい．そのため，採用検討対象となった格フレームが 2-C1 である可能性は考えないことにする．そのため，採用検討対象となった格フレームは全て採用する．

各格フレームが，どれに該当するかは，第 1 次格フレームでできた格フレームとの式 (2) で求められる確率値が最も高いものを選ぶ．

[*6] 「{ チップ, . . . } を弾む」のような他動詞パターンがあるが，ガ格とヲ格が交替しているわけではないので，自他両用とはいえるが自他同形ではない．

**第 3 次格フレーム**

第 3 次格フレームは以下のように分類できる．

3-A ガ格をヲ格に交替させた第 1 次格フレームが存在する（狭義の自他同形の自動詞パターン）

(16) { 会員, . . . } が {< 数量 >+人, . . . } に 拡大する
→ { 企業, . . . } が { 会員, . . . } を {< 数量 > 人, . . . } に 拡大する

3-B ガ格**以外**をヲ格に交替させた第 1 次格フレームが存在する（広義の自他同形の自動詞パターン）

(17) a. { 子供, . . . } が { プール, . . . } で 泳ぐ
→ { 子供, . . . } が { プール, . . . } を 走る
b. { 子供, . . . } が { 親, . . . } に 頼る
→ { 子供, . . . } が { 親, . . . } を 頼る [*7]

3-C1 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と考えられる

(18) a. { 彼, . . . } が { 実行, . . . } に 移す
→ { 彼, . . . } が { 計画, . . . } を { 実行, . . . } に 移す
b. { 彼, . . . } が { 部屋, . . . } に 飾る
→ { 彼, . . . } が { 花, . . . } を { 部屋, . . . } に 飾る
c. { 彼, . . . } が { 女性, . . . } に 贈る
→ { 彼, . . . } が { 花, . . . } を { 女性, . . . } に 贈る

3-C2 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とも，述語自体にヲ格相当のものが含意されている格パターンとも考えられる

(19) a. { 彼, . . . } が { 渋谷, . . . } で { 友達, . . . } と飲む
（「飲む」が「酒を」を含意しているとも考えられる）
b. { 私, . . . } が { 施設, . . . } に 寄付する
（「寄付する」が「金品を」を含意してい

[*7] 「{ 子育て, . . . } を { 親, . . . } に 頼る」のヲ格とは交替できない

るとも考えられる）

c. { 男, ...} が { ダッシュ, ...} で 走る
（「走る」が「道を」を含意しているとも考えられる）

3-E 2-D のガ格を省略した用例からなる格フレームと考えられる

(20) a. { 盛況, ...} に 終了する
→ { イベント, ...} が { 盛況を ...} に 終了する

b. { 麺, ...} に 絡む
→ { スープ, ...} が { 麺, ...} に 絡む

3-F 3-A, 3-B, 3-C1, 3-C2, 3-E のいずれでもない (第 1 次格フレームとも第 2 次格フレームとも対応付かない自動詞パターン)

(21) a. { 人, ...} が { 恋, ...} に 落ちる

b. { 人, ...} が { 視界, ...} から 消える

これらのうち，採用したくないものは 3-C1 と 3-E であるが，3-C1 と 3-C2 との区別を付けることは難しいため，採用検討対象となった格フレームが 3-C1 である可能性は考えないことにする．

各格フレームが，どれに該当するかは，第 1 次格フレームと第 2 次格フレームにできた格フレームとの式 (2) で求められる確率値が最も高いものを選ぶ．そして，3-E であると判断されたものは採用しないことにする．

## 5. 評価実験

### 5.1 実験内容

4 節で提案した格フレーム構築法の有効性を確認した．具体的には，以下の 20 述語のそれぞれ頻度上位 10 格フレームに対して，提案手法を用いた場合と用いなかった場合（ベースライン）の格フレームの妥当さをそれぞれ評価した．

- 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

妥当な格フレームは次の 2 つの条件を満たすものとする．

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

### 5.2 用例収集

web から収集した日本語文約 69 億文に対して，記号を多く含む文の削除等のフィルタリングを行い，約 40 億文を得た．そして，それらに対して形態素・構文解析を行い，約 4 万述語に対して解析誤りの可能性が低い述語項構造 [1] を抽出した．

形態素解析には JUMAN[*8]，構文解析には KNP[*9] を用いた．なお，JUMAN・KNP のいずれも修正等を加えた開発版 [*10] で，公開版より新しいものを使っている．

クラスタリング時間短縮のために，以下の初期格フレームの枝刈りを行った．

- 各初期格フレームから頻度 3 未満の項を削除
- 用例数が 10 未満の初期格フレームを削除
- 1 つしか格が無い初期格フレームを削除

### 5.3 初期格フレームのクラスタリング

初期格フレームのクラスタリングには，2.3 節で述べた Chinese Restaurant Process に基づく手法 [10] に次の変更を加えたものを用いた．

まず，$P(w \mid f_j)$ の定義に単語間類似度を用いる変更を加えた．[10] の定義では，単語 $w$ に似た単語が $f_j$ に存在する場合であっても，$w$ 自身が $f_j$ に存在しなければ，$count(f_j, w) = 0$ となってしまう．そこで，本稿では単語間類似度を用いて $count(f_j, w)$ を $count_{sim}(f_j, w)$ に置換して定義する．なお，$w$ が属する格を $c$ としたとき，$f_j$ のうち $c$ に属する項の集合を $f_j(c)$ とする．

$$P(w \mid f_j) = \frac{count_{sim}(f_j, w) + \beta}{\sum_{t \in V} count(f_j, t) + \mid V \mid \cdot \beta} \quad (4)$$

$$count_{sim}(f_j, w) = \sum_{x \in f_j(c)} sim(x, w) * count(f_j, x) \quad (5)$$

単語間類似度 $sim(x, w)$ には，word2vec[*11] [12] を web から取得した日本語 1 億文に対して，次元数を 500，negative examples を 5 として実行して求めたベクトルの内積を用いた．ただし，出現が 100 未満の語との類似度を計算する場合や内積が 0.2 未満の場合は類似度は 0 とした．さらに，$count_{sim}(f_j, w)$ の計算を高速化するため，実際には式 (5) を厳密に計算せずに，$f_j(c)$ の頻度が上位 5 つの項の類似度の重み付き平均と $count(f_j, w)$ の積で近似した．

なお，ハイパーパラメータは予備実験によって求めた $\alpha = 0.001, \beta = 1.0$ とした．

### 5.4 実験結果

ベースラインと提案手法を比較すると，平均格フレーム数は 39.5 から 46.3 に増えたものの，妥当な格フレームの割

[*8] http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN
[*9] http://nlp.ist.i.kyoto-u.ac.jp/?KNP
[*10] 近日公開予定である
[*11] https://code.google.com/p/word2vec/

| 評価 | 格フレーム | 格 | 用例 |
|---|---|---|---|
| ✗ | 当てられる:1 | ガ格 215 | 手 73, 手の平 20, 刃 18, 指 18, 唇 15, ··· |
| | | ヲ格 96 | 手 35, 聴診器 12, ボール 9, ナイフ 9, ローター 6, ··· |
| | | ニ格 1120 | 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 | 一部 5 |
| | | ニ格 664 | 費用 224, 資金 142, 返済 78, 支払い 62, 経費 61, 運営費 53, 活動費 27, ··· |
| ✗ | 拡大する:1 | ガ格 13 | おら 9，画像 4 |
| | | ヲ格 17294 | 写真 15725, 画像 944, イメージ 236, 文字 163，··· |
| | | ニ格 100 | ⟨ 数量 ⟩ サイズ 53, ⟨ 数量 ⟩ 倍 29,··· |
| ✗ | 拡大する:9 | ガ格 140 | 対象 130, 事業所 7, ⟨ 数量 ⟩ 区 3 |
| | | ヲ格 3060 | 対象 2077, 一部 405 対象者 223, ··· |
| | | ニ格 511 | サイズ 143, { 数量 }pix 52, ウインドー 33, 画像 17, ··· |

**表 3** ベースラインで構築した格フレームの例

| 評価 | 格フレーム | 格 | 用例 |
|---|---|---|---|
| ✓ | 当てられる:3 | ヲ格 158 | 手 128, 指 18, 人差し指 12 |
| | | ニ格 74 | 額 19, 頬 12, 唇 9, 口 8, 胸 6 |
| ✓ | 当てられる:5 | ガ格 225 | 手の平 61, 指 57, 唇 42, 人差し指 30, ナイフ 24, 指先 11 |
| | | ニ格 106 | 唇 32, 額 22, 首筋 16, 頬 15, 上 5, 背中 4, ··· |
| ✗ | 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,··· |
| ✓ | 拡大する:10 | ガ格 3 | ⟨ 数量 ⟩ 区 3 |
| | | ヲ格 2806 | 対象 2077, 対象者 223, 適用 190, 助成 159, 輪 74, 募集枠 39, ··· |
| | | ニ格 353 | ⟨ 数量 ⟩ 人 29, ⟨ 数量 ⟩ 年生 27, ⟨ 数量 ⟩ 回 14, 全域 13, 商品 12, 企業 12, ··· |
| ✓ | 拡大する:18 | ガ格 186 | 対象 176, 助成 10 |
| | | ニ格 10 | 資産 4, 業種 3, ⟨ 数量 ⟩ 回 3 |
| ✗ | 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, ··· |

**表 4** 提案手法で構築した格フレームの例

| 述語 | ベースライン | 提案手法 | 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 | 191 | +46 |

**表 5** 各述語頻度上位 10 格フレーム中の妥当な格フレーム数の変化

合は 72.5%(145/200) から 95.5%(191/200) に改善した（表 5）．ベースラインで構築した格フレームの例を表 3 に，提案手法で構築した格フレームの例を表 4 に，それぞれ示す．用例の列の数値は頻度を表す．

**提案手法で改善した例**

ベースラインでは「当てられる:1」のようにガ格にもヲ格にも「手」をとる格フレームができている．「当てられる」の弁別格順序は

ニ格 > ガ格 > ヲ格 > ...

である．そのため，

(22) a. 手を おでこに 当てられる
b. 手が おでこに 当てられる

といった用例は，ニ格が同じ「おでこ」なので，同じ初期格フレームに属すため，最終的にも同じ格フレームに属してしまう．一方，提案手法ではそれらは別の初期格フレームに属し，また，その 2 つの初期格フレームはそれぞれ第 1 次格フレームと第 2 次格フレームに属すので，「当てられる:3」と「当てられる:5」のように別の格フレームができた．

「拡大する」の弁別格順序は

ヲ格 > ニ格 > ガ格 > デ格 > ...

であるので，ベースラインと提案手法のいずれでも次の用例は異なる初期格フレームに属する．

(23) a. 対象を 一部で拡大します

b. 対象が 一部で拡大します

しかし，ベースラインではその 2 つの初期格フレームがクラスタリングにてマージされる可能性が有り，他の格が類似していることによって同じ格フレームになってしまうことを防げない．提案手法ではマージされる可能性が無いので，「拡大する:10」と「拡大する:18」のように別の格フレームができた．

**提案手法で改善できなかった例**

提案手法において妥当ではないと判断された格フレームは 9 つあった．その原因は，大まかには 2 つある．

1 つ目は，述語項構造抽出の誤りよるものである．「拡大する:1」に「おら」があるが，これは

(24) おらが村の健康茶減肥茶の写真を拡大する

という用例の固有名詞「おらが村の健康茶」から誤って抽出されたものである．これは，形態素解析器の語彙を強化することで対処できると考える．また，「拡大する:1」のガ格とヲ格の両方に「画像」があるが，

(25) 画像をクリックで画像が拡大します。

という用例から「画像が画像を拡大する」という項構造を抽出してしまっていた．これは，構文解析の確信度付与ルールを修正することで，抽出しないようにできると考える．

2 つ目は，ヲ格を省略した格フレームを削除できていないことによるものである．「使える:2」は 3-C2 と判断されて，最終的な格フレームに採用された．しかし，述語「使える」に対して対象物をとらないのは不自然である．これは，ヲ格が省略されているかどうかを判定する方法を改善することで対処できると考える．

## 6. おわりに

本稿では，用例を格の出現の仕方に基づいて 3 種類に分け，それぞれについて格フレームを構築してから最後に統合する格フレーム構築手法を提案した．また，格パターンを正しく認識するために必要な格フレームの述語単位の定義と可能動詞・サ変動詞の語彙の整備も行った．評価実験では，日本語ウェブテキスト 40 億文に対して提案手法を用いた格フレームを構築し，従来手法と比較して妥当な格フレームの割合が大きく改善したことを確認した．

今後は，本手法で構築した格フレームを用いて述語項構造解析を行う予定である．

**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：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：自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル，自然言語処理，Vol. 14, No. 4, pp. 67–81 (2007).
[4] Sasano, R., Kawahara, D. and Kurohashi, S.: The Effect of Corpus Size on Case Frame Acquisition for Predicate-Argument Structure Analysis, *IEICE TRANSACTIONS on Information and Systems*, Vol. E93-D, No. 6, pp. 1361–1368 (2010).
[5] Baker, C. F., Fillmore, C. J. and Lowe, J. B.: The Berkeley FrameNet Project, *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 86–90 (1998).
[6] Palmer, M., Gildea, D. and Kingsbury, P.: The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, Vol. 31, No. 1, pp. 71–106 (2005).
[7] Materna, J.: LDA-Frames: An unsupervised approach to generating semantic frames, *Proceedings of the 13th International Conference CICLing 2012, Part I, volume 7181 of Lecture Notes in Computer Science*, pp. 376–387 (2012).
[8] Materna, J.: Parameter Estimation for LDA-Frames, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 482–486 (2013).
[9] Aldous, D.: Exchangeability and related topics, *École d'Été de Probabilités de Saint-Flour XIII*, Springer Berlin Heidelberg, pp. 1–198 (1985).
[10] Kawahara, D., Peterson, D., Popescu, O. and Palmer, M.: Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 58–67 (2014).
[11] 益岡隆志，田窪行則：基礎日本語文法・改訂版，くろしお出版 (1992).
[12] Mikolov, T., Kai, C., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *Proceedings of Workshop at International Conference on Learning Representations* (2013).