言い換え用例に基づく機能動詞構文と動詞句の同義性計算

藤田 篤 † 佐藤 理史 † 泉 朋子 ‡ 今村 賢治 ‡ 菊井 玄一郎 ‡
† 名古屋大学大学院工学研究科
{fujita,ssato}@nuee.nagoya-u.ac.jp
‡ 日本電信電話株式会社 NTT サイバースペース研究所
{izumi.tomoko,imamura.kenji,kikui.genichiro}@lab.ntt.co.jp

1 はじめに

大規模な電子化テキストが安価に入手できるようになり，そこから言語に関する知識，常識，評判情報，予期せぬトラブル等の様々な知識を獲得・マイニングする研究が盛んに行われている．これらの知識は，一般に，テキストデータ中に出現する語や統語構造の断片に関する統計量に基づいて獲得される．その際，例 (1) のような同義の表現（**言い換え**）であっても，形態・統語的に異なれば，各々個別に統計量が計上される．

(1) a. 経験を持つ ⇔ 経験したことがある
b. 強まる ⇔ 強くなる
c. 魅力的だ ⇔ 魅力がある

知識獲得・マイニングをより高度化するには，言い換えを適切に処理する，すなわち，所与の 2 つの表現が同義であるか否かを判定する技術が必要である．

ひとくちに言い換えと言っても様々な現象があるが，同一視すべき言い換えの種類やその優先度（実現可能性，効果の多寡）は，想定する応用タスクや目的に依存する．我々は，近年の知識獲得・マイニング研究の関心が**事態（コト）**に向いていることを受け，事態の記述に用いられることの多い，述語句の同義性の計算手法について検討している [2, 3].

本稿では，例 (1a) のような，機能動詞構文と機能動詞を含まない動詞句の同義性計算手法について述べる．機能動詞構文は比較的頻繁に使用される構文であるため，応用タスクにおいても一定の効果が期待できる．

本研究に取り組むねらいは次の 2 点に集約される．

- 機能動詞が担う文法的機能の全体像を明らかにする
- 機能動詞構文と機能動詞を含まない単純な動詞句の同義性を計算するための包括的な資源 (cf. [7, 4, 1]), および頑健で高精度な手法を実現する

2 機能動詞構文と動詞句の言い換え

2.1 機能動詞構文の定義

村木 [6] は,「実質的な意味を名詞にあずけて，みずからはもっぱら文法的な機能を果たす動詞」を機能動詞と呼んでいる．しかしながら，この定義に照らしても，機能動詞を客観的に同定することはできない．

そこで本稿では，言い換えの可否という観点から機能動詞構文を定義する．すなわち，名詞 n が格助詞 c を介して動詞 v の格となっている動詞句のうち，例 (2) のように，n の動詞形 $v(n)$ を主辞とし，たかだか助動詞や副詞などの表現 f を付加した形の同義の動詞句が存在するものを，機能動詞構文と呼ぶ．

(2) 経験 を 持つ ⇔ 経験 したことがある
n c v $v(n)$ f

したがって，動詞 v が村木の言う機能動詞であっても，上記のような同義の動詞句が存在しない場合，本稿では，v を機能動詞とは呼ばない．

なお，便宜的に，次のように用語を定義しておく．

LVC パターン: 機能動詞構文を構成しうる $\langle c, v\rangle$
LVC 候補: LVC パターン $\langle c, v\rangle$ と動詞化可能な名詞（動作性名詞）n からなる動詞句 $\langle n, c, v\rangle$
正規形候補: LVC 候補の n の動詞形 $v(n)$ を主辞とし，たかだか助動詞・副詞等 f を付加した形の動詞句
正規形: 正規形候補のうち LVC 候補と同義のもの

2.2 機能動詞構文に対する正規形の認め方

すでに触れたように，すべての LVC 候補が機能動詞構文であるとは限らない．例 (3) に，同じ LVC パターン ⟨ を, 与える ⟩ を持つ 3 つの LVC 候補と，各々に対するいくつかの正規形候補を示す．

(3) a. 希望**を与える** ··· ⇏ 希望する，* 希望させる
b. 影響**を与える** ··· ⇔ 影響**する**，⇏ 影響させる
c. 感動**を与える** ··· ⇏ 感動する，⇔ 感動**させる**

(3a) の LVC 候補に対しては，いずれの候補も正規形ではないし，他にも正規形が存在しないため，機能動詞構文ではない．一方，(3b), (3c) の LVC 候補には正規形が存在するが，それらに含まれる表現 f は異なる．

本稿では，LVC 候補とその正規形候補が同義であるか否かを，文献 [4, 2] と同様，文脈とは独立に判断する (cf. [7, 1, 3]). したがって，例 (4) のように，1 つの LVC 候補に対して複数の正規形を認める．

(4) a. （抜き打ち）検査**を受ける** ⇔ 検査**される**
b. （胃の）検査**を受ける** ⇔ 検査**してもらう**

表 1: LVC 候補の抽出結果

文数	16,600,730
句点で終わる文の数	9,056,338
LVC 候補を抽出した文の数	452,213
抽出対象の文節対全体の異なり数	159,913
LVC 候補 $\langle n, c, v\rangle$ の異なり数	38,867
動作性名詞 n の異なり数	5,632
LVC パターン $\langle c, v\rangle$ の異なり数	160

3 機能動詞が担う機能の分析

同義性計算手法の検討に先立ち，機能動詞構文とその正規形の言い換え用例を作成し，機能動詞が担う文法的機能について調査した．

3.1 使用する資源

言い換え用例の作成には，次の 3 つの資源を用いた．

新聞コーパス: 毎日新聞記事データ 1991-2002 年版．

動作性名詞のリスト: 自動的に作成した表記と動詞形の対 15,598 件のリスト．内訳は次の通りである．

- サ変名詞 7,420 表記．IPADIC [1]の「名詞–サ変接続」(12,033 表記）のうち，新聞コーパス中に，動詞としての用例[2]が 1 例以上観察されるもの．
- 和動詞の連用形と基本形の対 8,178 表記．IPADIC の「動詞–自立」(14,620 表記）のうち，連用形が「名詞–一般」でもあるもの．サ変名詞と重複する 42 表記[3]は和動詞を優先する．

LVC パターンのリスト: 人間が作成した 165 表記のリスト．次の 2 つの資料から得たものの和集合である．

- 文献 [6] で例が示されている 143 表記．
- 文献 [1] で収集された 40 表記．

3.2 言い換え用例の作成

言い換え用例の具体的な作成手順と結果を示す．

Step 1. 新聞コーパスから，動作性名詞のリストに含まれる n と LVC パターンのリストに含まれる $\langle c, v\rangle$ で構成される LVC 候補 $\langle n, c, v\rangle$ を抽出した．文末の 1 つ前の n と c で終わる文節と，v で始まり句点で終わる文末文節の対のみを対象として LVC 候補を抽出した結果を表 1 に示す．

Step 2. LVC パターンごとに LVC 候補をサンプリングした．頻度 10 以上，かつ $\langle c, v\rangle$ ごとに頻度の上位 10 位に入るもの抽出した結果，140 種類の LVC パターンに対する 1,095 件の LVC 候補を得た．

Step 3. 各 LVC 候補に対し，図 1 の決定木に従って正規形を付与した．結果の分布を表 2 に示す．正規形を付与する際に用いられた助動詞・副詞等の表現は，図 2 に示す 26 種類であった．

[1] http://chasen.naist.jp/hiki/ChaSen/, ver.2.6.3

[2] 「する」,「し」,「させ」,「され」,「でき」を後続．

[3] 「引っ越し」,「割り引き」など．

$\langle n, c, v\rangle$**に対して正規形が存在するか:**

- 存在する $\cdots$ n **を動詞化するだけで正規形となるか:**
  - 正規形となる $\cdots$ 【ϕ】
  - 助動詞・副詞等が必要 $\cdots$ 【助動詞等/副詞等】
- 存在しない $\cdots$ n **は動詞化できたか:**
  - できない $\cdots$ 【n が動詞化不可】
  - できる $\cdots$ **正規形がないのはなぜか:**
    - 意味を保存できないため $\cdots$ 【言い換え不可】
    - 文法的に不適格になるため $\cdots$ 【言い換え不可】

図 1: 言い換え付与作業の決定木

表 2: 言い換え用例における助動詞・副詞等の分布

分類	$\langle n, c, v\rangle$の異なり数
n が動詞化不可	56
言い換え不可	174
ϕ	377
助動詞等	404
副詞等	74
副詞等+助動詞等	10
合計	1,095

3.3 言い換え用例の分析

表 1 に示すように，約 5%の文の文末から LVC 候補が抽出された．動詞化できない動作性名詞[4]を含む候補 (56 件，36 語）や (3a) のように正規形を与えられない（機能動詞構文ではない）候補（174 件）が含まれるものの，機能動詞構文の出現確率は決して低くない．

正規形を与える際に付加された表現 f の大半は，表 2 に示すように，何もない（“ϕ”）か，助動詞程度であった．ただし，図 2 に示すように，助動詞の組み合わせや，副詞等を用いなければならない場合もあった．

得られた言い換え用例を LVC パターン $\langle c, v\rangle$ ごとにみると，新たに次の 2 点が明らかになった．

- 村木 [6] が示したように，確かにその意味の希薄さも認められるものの，助動詞・副詞等を用いても正規形を与えられない $\langle c, v\rangle$ が 6 種類あった[5]．
- 例 (3) のように，同じ $\langle c, v\rangle$ を持つ LVC 候補でも，動作性名詞 n によって，機能動詞構文ではなかったり，正規形に含まれる表現 f が異なる例があった．同じ $\langle c, v\rangle$，同じ f を持つ正規形が与えられた動作性名詞の間には類似性が見られるものの，その観点は，自動詞や他動詞という統語的性質から，非対格性やアスペクト的特徴などの語彙意味的性質，語の意味そのものまで多様である．

4 用例に基づく同義性の計算

LVC 候補 $\langle n, c, v\rangle$ と正規形候補 $\langle v(n), f\rangle$ の同義性を計算するには，各 LVC パターン $\langle c, v\rangle$ に対して，$\langle v(n), f\rangle$ を正規形とするような動作性名詞 n の選択制限を記述する必要がある．しかしながら，用例を見る限り，n の選択制限を既存の資源に照らして記述す

[4] 「施設」,「判決」,「意思」など．

[5] 〈を, 打ち切る〉，〈を, やめる〉，〈に, 進む〉など．

助動詞等–ヴォイス (3): 「させる」,「される」,「してもらう」
助動詞等–アスペクト (5): 「し始める」,「し続ける」,「している」,「したことがある」,「し終わる」
助動詞等–ムード (4): 「してしまう」,「しなければならなくなる」,「できる」,「しようとする」
助動詞等–ヴォイス+アスペクト (1): 「されている」
助動詞等–ヴォイス+ムード (1): 「させようとする」
副詞等 (9): 「繰り返し」,「あれこれと」,「十分に」,「頻繁に」,「一層」,「急いで」,「一所懸命」,「互いに」,「より深く」
副詞等+助動詞等 (3): 「よく～することになる」,「よく～されることになる」,「うまく～しようとする」

図 2: 言い換え用例に用いられた助動詞・副詞等の一覧

f **が一意に決まるか:**
- 決まらない $\cdots$ **言い換え不可の用例はあるか:**
  - ある $\cdots$ 【各 f に対して Class/Class(NG)】
  - ない $\cdots$ 【各 f に対して Class】
- 決まる $\cdots$ f =**"NG":**
  - f ="NG" $\cdots$ 【Any(NG)】
  - $f \neq$"NG" $\cdots$ **用例の数は生産性を期待できるほどあるか:**
    - 用例数 $\geq 5 \cdots$ 【Any】
    - 用例数 $< 5 \cdots$ 【Instance】

図 3: $\langle c, v \rangle$ ごとの同義性判定規則群のタイプの決定木

ることは困難である．そこで，我々は，選択制限を満足するか否か，しいては $\langle n, c, v \rangle$ と $\langle v(n), f \rangle$ の同義性を，3 節で得た用例との類似度に基づいて判定する．

4.1 同義性判定規則

"ϕ" や "言い換え不可 (NG)" も f の一種とみなし，$\langle c, v, f \rangle$ ごとに次の形式の規則を 1 つ作成する．

$$\langle Type, S_n, c, v, f \rangle$$

S_n は動作性名詞 n の用例集合である．ただし，その解釈は，下記の通り，規則のタイプ $Type$ に応じて異なる．

Any: $\langle c, v \rangle$ に対して，あらゆる LVC 候補が $\langle v(n), f \rangle$ と同義．すなわち，S_n は無視する．

Class: $\langle c, v \rangle$ に対して，S_n 中の動作性名詞の用例と類似する n を持つ LVC 候補のみが，$\langle v(n), f \rangle$ と同義．すなわち，S_n は典型例と解釈する．

Instance: $\langle c, v \rangle$ に対して，特定の LVC 候補のみ，$\langle v(n), f \rangle$ と同義．すなわち，S_n は厳密な語彙的制約と解釈する．

3 節で作成した 1,095 件の言い換え用例から，$\langle c, v \rangle$ ごとに図 3 の決定木に従って 1 つ以上の同義性判定規則を作成した．ここでは，各 $\langle c, v \rangle$ に対する規則群の $Type$ はすべて同じとしている．作成した 233 件の規則の内訳を表 3 に，例を図 4 に示す．$Type$ ="Instance", f ="NG" という規則は，動詞化できない動作性名詞集合 X に対する例外規則 $\langle$ Instance, X, $*$, $*$, NG $\rangle$ である．

4.2 同義性判定アルゴリズム

LVC 候補 $\langle n, c, v \rangle$ と正規形候補 $\langle v(n), f \rangle$ の同義性は，図 5 の決定木に従って判定する．

例 (4) のように，1 つの LVC に対して複数の正規形が存在する場合がある．したがって，f に曖昧性がある

表 3: 同義性判定規則の内訳

$Type$	$\langle c, v \rangle$ の数	$f \neq$"NG"	$f =$"NG"
Any	48	42	6
Class	72	111	53
Instance	20	20	1
合計	140	173	60

$\langle Type, Sn, c, v, f \rangle$
$\langle$ Instance, { 注意, 努力, 長考 }, を, 払う, する $\rangle$
$\langle$ Instance, { 努力 }, を, 傾ける, する $\rangle$
$\langle$ Any, $*$, を, 行う, する $\rangle$
({ 試合, 調査, 活動, 会談, 協議, 演説 })
$\langle$ Any, $*$, が, 目立つ, 頻繁に $\rangle$
({ 動き, 活躍, 発言, 意見, 落ち込み, ミス })
$\langle$ Any, $*$, を, 打ち切る, NG $\rangle$
({ 運転, 捜索, 契約, 会見, 調査, 捜査 })
$\langle$ Class, { 影響, 刺激, 評価, 許可, 示唆 }, を, 与える, する $\rangle$
$\langle$ Class, { 感動, 感銘, 安らぎ }, を, 与える, させる $\rangle$
$\langle$ Class, { 希望 }, を, 与える, NG $\rangle$

図 4: LVC 候補と正規形候補の同義性判定規則の例

$\langle c, v, f \rangle$ **に関する規則が存在するか:**
- 存在しない $\cdots$ 【$\neq$】
- 存在する $\cdots$ **規則のタイプは何か:**
  - $Type$ ="Any" $\cdots$ 【=】
  - $Type$ ="Instance" $\cdots$ **規則の適用条件を満たすか:**
    - 満たす．すなわち，$n \in S_n \cdots$ 【=】
    - 満たさない $\cdots$ 【$\neq$】
  - $Type$ ="Class" $\cdots$ **規則の適用条件を満たすか:**
    - 満たす．すなわち，$\langle c, v \rangle$ を共有する複数の規則の中で，当該規則の $Sim_{rule}(n, S_n)$ が最大 $\cdots$ 【=】
    - 満たさない $\cdots$ 【$\neq$】

図 5: 同義性判定の決定木

（$Type$ ="Class" の）$\langle c, v \rangle$ については，個々の f に対して個別に同義性を判定することが妥当と考えられる．しかしながら，3 節で述べたように，動作性名詞 n の間の類似性の観点は様々であるため，全ての $\langle c, v \rangle$ に共通の閾値を決めることは容易ではない．そこで，用例集合との類似度が最大となる規則のみを用いる．

LVC 候補の動作性名詞 n と $Type$ ="Class" なる規則の用例集合 S_n との類似度は，次式で算出する．

$$Sim_{rule}(n, S_n) = \max_{n_i \in S_n} Sim_{verb}\big(v(n), v(n_i)\big).$$

ここでは，n および $n_i \in S_n$ の動詞形の類似度を用いている．2 つの動詞間の類似度は次式で与える．

$$Sim_{verb}(v_1, v_2) = 1/DS_{JS}\big(P(Z|v_1), P(Z|v_2)\big).$$

ここで，DS_{JS} は確率分布間の Jensen-Shannon divergence [5] である．個々の動詞に対する確率分布 $P(Z|v)$ は次の手順で学習した．

Step 1. 新聞コーパスから，名詞 n が格助詞 c を介して動詞 v の格となっている動詞句 $\langle n, c, v \rangle$ を抽出した．

Step 2. 頻度 2 以上の $\langle n, c, v \rangle$ から動詞 v と格要素 $\langle n, c \rangle$ の共起頻度行列を作成し，PLSI 学習パッケージ[6]を用いて各動詞 v の，各隠れ変数 $z \in Z$ への帰属確率 $P(z|v)$ を推定した．今回は，隠れ変数の数 $|Z|$ を適当に 1,000 とした．

[6] http://chasen.org/~taku/software/plsi/

表 4: 動詞間の類似度モデルの諸元

文数	16,600,730
抽出した $\langle n, c, v\rangle$ ののべ数	28,270,464
抽出した $\langle n, c, v\rangle$ の異なり数	5,624,446
モデル構築に用いた $\langle n, c, v\rangle$ ののべ数	24,791,471
モデル構築に用いた $\langle n, c, v\rangle$ の異なり数	2,145,453
v の種類	12,231

このようにして得られたモデルの諸元を表 4 に示す．

5 同義性判定実験

文献 [4, 2, 3] では自動生成した表現の精度を評価した．これに対し，我々は，実際に存在する表現を正規形候補としてあらかじめ正解データを作成した．そして，手法全体の性能を，同義とすべき候補を検出するタスクの再現率，精度等によって評価した．

5.1 正解データの作成

正解データは，次の手順で作成した．

Step 1. 表 1 の LVC 候補 $\langle n, c, v\rangle$ から，次の 3 つの要件をすべて満たす 4,988 件を抽出した．

- 頻度 10 以上
- 言い換え用例としてサンプリングしていない
- $\langle c, v\rangle$ に対する同義性判定規則が存在する

Step 2. 各 $\langle n, c, v\rangle$ の頻度を考慮して，4,988 件からランダムに 200 件サンプリングした．この 200 件に対する $\langle c, v\rangle$ の異なり数は 38 であった．

Step 3. 各 $\langle n, c, v\rangle$ に対する正規形候補 $\langle v(n), f\rangle$ を新聞コーパスから抽出した．LVC 候補と同様，文末の文節を対象とし，$v(n)$ で始まり句点で終わるものを抽出した．ただし，f は図 2 に含まれる表現に限定した．この結果，$\langle n, c, v, v(n), f\rangle$ を 1,397 件得た．

Step 4. 各 $\langle n, c, v, v(n), f\rangle$ の同義性を人間が判定した結果，同義である (“⇔”) 197 件と，同義でない (“⇏”) 1,200 件を得た．

5.2 同義性判定の結果と考察

提案手法を用いて上述の 1,397 件の同義性を判定した結果を表 5 に，その再現率，精度等を表 6 に示す．表 6 における BOW.Lin〜MOD.skew の 4 行は，2 つの表現対が言い換えであるか否かを各表現と共起する表現の分布の比較に基づいて計算するという手法 [2] を同じデータに適用した結果である[7]．提案手法は，既存の手法よりも再現率が若干劣るものの，それ以外の指標では既存の手法を上回る性能を示した．

102 件の認定もれのうち 71 件は規則の不足による．3 節で作成した用例だけでは，個々の $\langle c, v\rangle$ に対する f を網羅できていないことが原因であった．残りの 31 件はすべて，$Type$ =“Class” という規則が適用された LVC 候補であった．新規の f を発見した場合は規則を，$Type$ =“Class” の規則で誤認定が生じた場合は用例を追加すれば規則全体の質は改善できる．この際，$Type$ =“Class” の規則が増えるため，類似度の尺度がより重要になる．一方，泉ら [3] が指摘するように，目的によっては，同義でなく含意関係も捉える必要がある．ゆえに，f の粒度に関する検討も必要である．

200 件の LVC 候補のうち，2 種類以上（最大 2 種類）の正規形を持つものが 23 件あった．唯一の正規形しか認めない現状のアルゴリズムの修正も必要である．

[7] 2009 年 1 月 6 日に収集したウェブページのスニペットを使用．

表 5: 提案手法による同義性判定結果

		出力					
		全体		Any 分		Class 分	
		⇔	⇏	⇔	⇏	⇔	⇏
正解	⇔	95	102	65	47	30	55
	⇏	27	1,173	10	649	17	524

表 6: 各手法の性能

モデル	再現率	精度	F 値	正解率
すべて ≠	-	-	-	.859
BOW.Lin	.34	.35	.35	.816
BOW.skew	.36	.36	.36	.820
MOD.Lin	.50	.50	.50	.859
MOD.skew	**.52**	.53	.52	.866
提案手法	.48	**.78**	**.60**	**.908**
Any 分	.58	.87	.70	.926
Class 分	.35	.64	.45	.885

6 おわりに

本稿では，先行研究で扱っていたよりも幅広い種類の機能動詞，およびそれらが表す機能に対象を広げ，機能動詞構文が表す文法的機能が，機能動詞を含まない単純な動詞句ではどのような表現によって担保されるかを調査した．そして，分析のために作成した用例に基づいて，与えられた表現が機能動詞構文であるか否かを同定，機能の曖昧性を解消，そして同義性を計算する手法を提案し，その有効性を検証した．

参 考 文 献

[1] 藤田篤, 降幡建太郎, 乾健太郎, 松本裕治. 語彙概念構造に基づく言い換え生成—機能動詞構文の言い換えを例題に. 情報処理学会論文誌, Vol. 47, No. 6, pp. 1963–1975, 2006.

[2] A. Fujita and S. Sato. A probabilistic model for measuring grammaticality and similarity of automatically generated paraphrases of predicate phrases. In *Proceedings of the 22nd International Conference on Computational Linguistics* (*COLING*), pp. 225–232, 2008.

[3] 泉朋子, 今村賢治, 菊井玄一郎, 藤田篤, 佐藤理史. 正規化を指向した機能動詞表現の述部言い換え. 2009. (in this proceedings).

[4] 鍛治伸裕, 黒橋禎夫. 迂言表現と重複表現の認識と言い換え. 自然言語処理, Vol. 11, No. 1, pp. 81–106, 2004.

[5] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, Vol. 37, No. 1, pp. 145–151, 1991.

[6] 村木新次郎. 日本語動詞の諸相. ひつじ書房, 1991.

[7] 奥雅博. 日本文解析における述語相当の慣用的表現の扱い. 情報処理学会論文誌, Vol. 31, No. 12, pp. 1727–1734, 1990.