語構造情報を用いた日本語複合動詞の言い換え

野口 真人　　梶原 智之　　小町 守
首都大学東京
{noguchi-mahito, kajiwara-tomoyuki}@ed.tmu.ac.jp, komachi@tmu.ac.jp

1 はじめに

言語処理において，語彙の換言は重要な役割を果たす．例えば，機械翻訳の際に言い換えを行うことで，より自然な訳文を生成できることが知られている [1]．また，機械翻訳の評価においても言い換えの処理を行うことで，より正確な評価が可能となることが知られている [2]．複雑な語の換言は，子ども，高齢者，外国人などの言語学習者，障害者に対して言語能力の差を埋めるための手段としても有用である [3]．

これまで日本語では，名詞 [4] や動詞 [5] などの内容語の言い換え [6, 3] が研究されている．

梶原法 2013 [6] は，語釈文を用いた内容語の言い換え手法である．語釈文から見出し語と同じ品詞の単語を抽出し，主要部終端型である日本語の特徴を利用して，末尾に近い単語を言い換えとする．

梶原法 2015 [3] は，前節で紹介した語釈文を用いた内容語の言い換え手法を改良している．語釈文から見出し語と同じ品詞の単語を複数抽出し，最も尤もらしい単語を言い換えとして獲得する．一方で，複合動詞の言い換えについての研究は未だ行われていない．

そこで本稿では，神崎 [7] が作成した「複合動詞レキシコン」[1] を用いて，複合動詞の言い換えを行う．複合動詞レキシコンは，複合動詞に関して質・量ともに充実したデータベースであり，複合動詞 2,756 語に関する情報が収録されている．収録されている内容は複合動詞の表記，定義文，格情報，例文，語構造，前項動詞，後項動詞である（図 1）．本研究ではこのうち語構造，前項動詞，後項動詞を利用した体系的な言い換えを行い，他の手法 [3, 6] を用いた言い換えよりも 5.5 ポイント高い 56.8%の正解率を示した．また，機械学習を用いた言い換えの推定を行い，素性として語構造情報を用いた場合，他の素性の場合よりも高い正解率を示した．

[1] http://vvlexicon.ninjal.ac.jp/

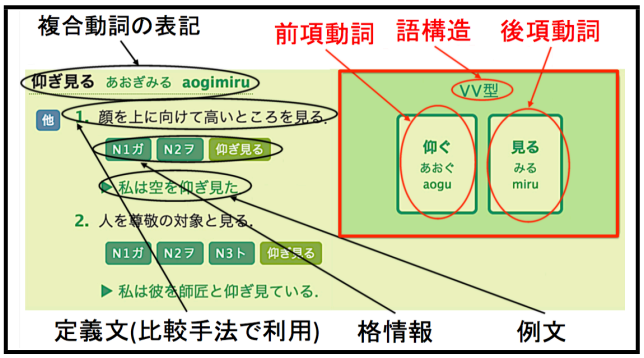

図 1: 複合動詞レキシコンに収録されている情報

表 1: 複合動詞レキシコン内におけるそれぞれの語構造の割合

	VV 型	Vs 型	pV 型	V 型
数	1,651 語	833 語	169 語	116 語
割合	59.6%	30.1%	6.1%	4.2%

2 語構造情報を用いた言い換え

本研究では，複合動詞を別の動詞に言い換える．以下にルールベースの手法（提案手法 1）と機械学習ベースの手法（提案手法 2）についてそれぞれ記述する．

2.1 ルールベースの手法（提案手法 1）

語構造情報を用いたルールベースの言い換えでは，前項動詞もしくは後項動詞のどちらかに言い換えを行う．どちらの動詞に言い換えるかは，語構造（VV 型，Vs 型，pV 型，V 型）により決定する．

表 1 に，複合動詞レキシコン内におけるそれぞれの語構造の数と割合を示す．これらの語構造は複合動詞の意味的な機能を大まかに示すため，この情報を言い換えに利用する．本節では，語構造を用いた言い換えの方法を示す．表 2 に，複合動詞の語構造による言い換え例を示す．

VV 型：動詞＋動詞　VV 型は，自立した動詞同士の組み合わせであり，「V1 して V2 する」「V1 しつつ V2

表 2: 複合動詞の語構造による言い換え例

見出し語	語構造	例文	方法
仰ぎ見る	VV 型	私は空を｛仰ぎ見た → ×仰いだ ○見た｝．	後項動詞に言い換え
ぶち壊す	pV 型	彼はドアを｛ぶち壊し → ×ぶっ ○壊し｝て侵入してきた．	
ざわめき立つ	Vs 型	全校生徒が｛ざわめき立っ → ○ざわめい ×立っ｝た．	前項動詞に言い換え
入れ込む	VV 型	シャツをズボンに｛入れ込んだ → ○入れた ×込んだ｝．	
植え込む	VV 型	庭に植木を｛植え込んだ → ○植えた ×込んだ｝．	
掛け合う	V 型	彼は店員に値引きを｛掛け合っ → ×掛け ×合っ｝た．	言い換えない

する」などの意味を持つ複合動詞である．日本語は主要部終端型の言語であるため，VV 型の複合動詞は後項動詞に言い換える．

pV 型：接頭辞化した動詞＋動詞　pV 型は，接頭辞化した動詞と動詞の組み合わせである．前項動詞の意味は希薄になっているため，pV 型は後項動詞に言い換える．

Vs 型：動詞＋補助的な動詞　Vs 型は，自立した動詞と補助的な動詞の組み合わせである．後項動詞の意味は希薄になっているため，Vs 型は前項動詞に言い換える．

VV 型は後項動詞に言い換えると先に述べたが，後項動詞が「込む」の複合動詞は,「込む」と言い換えるのは不適切である．よって「〜込む」という複合動詞は，Vs 型に再分類することで，適切な言い換えを行うことができる（表 2 の「入れ込む」・「植え込む」）．

V 型：一般化　V 型は，一般化され 1 語として固定化された複合動詞である．提案手法 1 では V 型の複合動詞を言い換えることはできないが，V 型は全体の 4%しか存在していないため，無視しても大きな影響はないので，無視する．

2.2　機械学習ベースの手法（提案手法 2）

機械学習ベースの手法による言い換え獲得では，それぞれの文中に存在する複合動詞が，前項動詞に言い換えられるか否かの 2 値の判定を行った．同じく，後項動詞やその他の動詞においても言い換えられるか否かの 2 値の判定をそれぞれ行った．学習に用いる素性を以下に示す．

- 単語 bag-of-words
- 複合動詞の語構造情報
- 複合動詞の前後 3 単語を文脈窓とした文脈ベクトル

3　語構造による言い換えの評価実験

本節では，まず 3.1 節で実験設定について紹介する．3.2 節では評価方法を記述する．3.3 節では言い換えの正解率を先行研究と比較し，提案手法の優位性を示す．

3.1　実験設定

本節では，比較手法および提案手法 2 の実験設定を紹介する．以下にそれぞれの手法の実験設定を記述する．

3.1.1　辞書の語釈文を用いた言い換え（梶原法 2013, 2015）

本実験では複合動詞レキシコンの定義文を MeCab（ver0.996，IPAdic ver2.7.0）[2] で解析した．

梶原法 2013 では，語釈文の中で最も末尾に近い動詞を複合動詞の言い換えとする．

梶原法 2015 では，シソーラスを用いて語釈文中から尤もらしい単語を選んで複合動詞の言い換えとしていたが，例えば日本語 WordNet [8]（Ver.1.1）には多くの複合動詞が収録されていない．そこで本研究では，word2vec [9] を用いて尤もらしい単語を選ぶ．日本語 Wikipedia（2015 年 06 月 05 日アクセス，5.07GB）[3] を，複合動詞 2,756 語をユーザ辞書 [4] として追加した MeCab で分かち書きをして，word2vec の学習を行う．その後それぞれの複合動詞と類似度が高い上位 100 単語までを参照し，その中に定義文から MeCab で抽出した動詞が存在する場合，その単語を言い換えとして獲得する．上位 100 位までに抽出した動詞が存在しない場合は，梶原法 2013 と同じように末尾に近い動詞を言い換えとする．

[2] http://taku910.github.io/mecab/

[3] http://dumps.wikimedia.org/jawiki/latest/

[4] 複合動詞のコストを-15000 とすることで，必ず複合動詞が選択されるようにした．また文脈 ID は，左文脈 ID を前項動詞の左文脈 ID から，右文脈 ID を後項動詞の右文脈 ID からそれぞれ用いることで，前後の語を正しく認識できるようにした．

表 3: 手法ごとの言い換え正解率

	提案手法 1	梶原法 2015	梶原法 2013
正解率	**56.8%**	51.3%	50.3%

複合動詞レキシコンから抽出した例文 300 文を開発データとして word2vec のパラメータ調整を行ったところ，次元 50，文脈窓 3，CBOW，階層的ソフトマックスあり，ネガティブサンプリング 10 の結果が最高であったため，このパラメータを本実験で利用した．

3.1.2 機械学習を用いた言い換え

機械学習を用いた言い換えには，Support Vector Machine（SVM）を利用した．分類器には scikit-learn[5] の RBF カーネル SVC を使用した．正則化パラメータは，グリッドサーチを行いペナルティパラメータ $C = 100$ とした．また，3.1.1 節と同じ設定で文を分かち書きした．

3.2 評価方法

まず，提案手法 1，梶原法 2013 および梶原法 2015 を用いて言い換えの収集を行った．BCCWJ [10] から無作為に抽出した 1,000 文（複合動詞 1 語につきそれぞれ 10 文× 100 語）の言い換えを行い，文脈を考慮した言い換えを人手で評価した．評価は日本語を母語とする工学系学部生 3 人が行った．一致率は $\kappa = 0.68$ であり十分一致しているといえる．

また，それぞれの文において，複合動詞が前項動詞に言い換えられるかを 2 値でラベル付けした．同じように，後項動詞やその他の動詞に言い換えられるかもそれぞれ 2 値でラベル付けした．評価実験では，複合動詞が前項動詞，後項動詞またはその他の動詞に言い換えられるかどうかをそれぞれ予測する．評価方法については，テストデータと同じ複合動詞の文が入らないようにした学習用データを用意し，10 分割交差検証を行い，正解率，適合率，再現率および F 値をそれぞれ求めた．

3.3 実験結果

表 3 に，提案手法 1，梶原法 2013 および梶原法 2015 で獲得した複合動詞の言い換えの正解率を示す．提案手法 1 が一番高い正解率を示した．

表 4 に，提案手法 2 で素性を変更して実験した結果を示す．前項動詞・後項動詞の予測においては，提案手法 1 を用いた場合と語構造情報のみを用いて学習した場合，最も良い結果となった．この結果は，語構造情報が言い換えの予測において重要な素性であることを示している．その他の動詞の予測においては，正解率と適合率は語構造情報を用いて学習した場合，最も良い結果となった．再現率と F 値は bag-of-words を用いて学習した場合，最も良い結果となった．この結果は，その他の動詞に言い換えられるかどうかの予測においては，語構造情報が必ずしも重要な素性ではないことを示している．

4 考察

表 3 の実験結果は，複合動詞レキシコンが複合動詞の言い換え獲得に有効であり，特に語構造を用いた言い換えの獲得が効果的であることを示している．表 5 に提案手法による言い換えの例を，表 6 に語構造ごとの提案手法における言い換えの正解率をそれぞれ示す．前述した通り提案手法では V 型の言い換えは行わないため，正解率は 0%となっている．正解率の最も高い語構造は pV 型であった．

次に，言い換えできなかった例について考察を行う．

VV 型においては，不正解となった文は 255 文存在する．そのうち，前項動詞にも後項動詞にも言い換えが不可能な文が，163 文（63.9%）存在した．例えば表 5 の「見落とす」は,「見る」とも「落とす」とも言い換えられない．また，VV 型でも前項動詞に言い換えられるような文が，92 文（36.1%）存在した．例えば表 5 の「考え出す」は,「出す」とは言い換えられないが,「考える」とは言い換えられる．

Vs 型においては，不正解となった文は 127 文存在し，そのうち 112 文（88.2%）は一般化された複合動詞であった．例えば表 5 の「思い切る」は,「思う」にも「切る」にも言い換えられず，1 語となっていることがわかる．また，一般化されてはいないが後項動詞の要素を無視できない文が 15 文（11.8%）存在した．例えば表 5 の「出し渋る」は，1 語とはなっていないが「渋る」は否定の意味を含んでおり，前項動詞のみに言い換えるのは不適切である．

pV 型においては，不正解とされた文が 25 文存在し，その全てが前項動詞にも後項動詞にも言い換えられない一般化された複合動詞であった．例えば表 5 の「取り巻く」や「取り組む」は，前項動詞にも後項動詞にも言い換えられない．

また，提案手法が正解となった場合でも，提案手法よりも語釈文を用いた言い換えの方がより流暢な言い

[5] http://scikit-learn.org/stable/index.html

表 4: 機械学習ベースでの言い換え結果（左から前項動詞，後項動詞，その他の動詞）

	正解率	適合率	再現率	F 値
ALL	.588 / .699 / .638	.538 / .615 / .335	.516 / .632 / **.186**	.526 / .623 / **.239**
ALL - BoW	.618 / .718 / .692	.569 / .612 / .490	**.574**/ .777 / .157	.572 / .685 / .238
ALL - 語構造	.519 / .547 / .649	.452 / .405 / .347	.396 / .317 / .167	.423 / .356 / .225
ALL - 文脈	.582 / .714 / .633	.531 / **.629**/ .281	.509 / .670 / .127	.520 / .649 / .175
bag-of-words	.507 / .549 / .637	.434 / .397 / .279	.363 / .279 / .118	.395 / .328 / .166
語構造情報	**.662**/ **.740**/ **.696**	**.647**/ .616 / **.520**	.525 / **.906**/ .085	**.580**/ **.733 / .146**
文脈ベクトル	.525 / .587 / .684	.458 / .460 / .396	.381 / .274 / .062	.416 / .343 / .107
ルールベース	**.662**/ **.740**/ —	**.647**/ .616 / —	.525 / **.906**/ —	**.580**/ **.733**/ —

表 5: ルールベースによる言い換えの例

見出し語	語構造	例文
取り除く	VV 型	歯茎を｛取り除け → ○除け｝ば，症例は回復する．
見つけ出す	Vs 型	アイデアを｛見つけ出し → ○見つけ｝てプロットする．
引き渡す	pV 型	その女の子を我々に｛引き渡さ → ○渡さ｝なければならない．
見落とす	VV 型	何回か信号を｛見落とし → ×落とし｝ そうになった．
考え出す	VV 型	捨て身に近い反復攻撃を｛考え出し → ×出し｝ たのは彼である．
思い切る	Vs 型	彼は現状を｛思い切っ → ×思っ｝ て変えた．
出し渋る	Vs 型	彼はお金を｛出し渋っ → ×渋っ｝ た．
取り巻く	pV 型	物流を｛取り巻く → ×巻く｝ 環境は大きく変わった．
取り組む	pV 型	彼は住宅の問題に｛取り組ん → ×組ん｝ でいる．
言い切る	Vs 型	彼ははっきりとそう｛言い切っ → ○言っ｝た．

表 6: ルールベースにおける語構造ごとの言い換えの正解率

	VV 型	Vs 型	pV 型	V 型
正解率	57.2%	60.3%	73.0%	0.0%

換えを獲得できた複合動詞もある．例えば，表 5 の「言い切る」は「言う」と言い換えても正解であるが，「断言する」と言い換える方がより流暢である．

5 おわりに

本研究では，複合動詞レキシコンの語構造情報を用いて，複合動詞を前項動詞または後項動詞に言い換えた．我々の主張は以下の通りである．

- 複合動詞の言い換え獲得には，語構造を用いた言い換えの獲得が最も効果的である．

今後は，3.2 節で紹介したデータセットについて公開する予定である．

参考文献

[1] Wei He, Hua Wu, Haifeng Wang, and Ting Liu. Improve SMT quality with automatically extracted paraphrase rules. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 979–987, 2012.

[2] Hiroshi Kanayama. Paraphrasing rules for automatic evaluation of translation into Japanese. In *Proceedings of the second international workshop on Paraphrasing*, pp. 88–93, 2003.

[3] 梶原智之, 山本和英. 語釈文を用いた小学生のための語彙平易化. 情報処理学会論文誌, Vol. 56, No. 3, pp. 983–992, 2015.

[4] 藤田篤. 語釈文を利用した普通名詞の同概念語への言い換え. 言語処理学会 第 7 回年次大会, 2001.

[5] 鍜治伸裕, 河原大輔, 黒橋禎夫, 佐藤理史. 格フレームの対応付けに基づく用言の言い換え. 自然言語処理, Vol. 10, No. 4, pp. 65–81, 2003.

[6] 梶原智之, 山本和英. 小学生の読解支援に向けた複数の換言知識を併用した語彙平易化と評価. 言語処理学会 19 回年次大会発表論文集, pp. 272–275, 2013.

[7] 神崎享子. 『複合動詞レキシコン』ver.1 —形態的・統語的・意味的情報付与—. 言語処理学会第 19 回年次大会発表論文集, pp. 761–764, 2013.

[8] Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. Japanese semcor: A sense-tagged corpus of japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pp. 56–63, 2012.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of International Conference on Learning Representations*, 2013.

[10] 国立国語研究所コーパス開発センター. 『現代日本語書き言葉均衡コーパス』利用の手引，第 1.0 版. 2011.