述語句の形態・構文的言い換えの確率的生成モデル*

加藤 修平 † 藤田 篤 † 佐藤 理史 †
† 名古屋大学大学院工学研究科
shuhei@sslab.nuee.nagoya-u.ac.jp, {fujita,ssato}@nuee.nagoya-u.ac.jp

1 はじめに

自然言語処理の諸タスクにおいて，同じ意味を表す異なる言語表現（言い換え）を処理する技術の必要性が論じられている [2]．これまでに，(1) に示すような同義表現対・統語構造変換パターンの記述，それらの自動獲得といった，言い換え知識の整備に主眼を置いた研究が多くなされてきた．

(1) a. 横領する ⇔ 着服する ⇔ 使い込む
b. $N:C:V \Rightarrow adv(V):vp(N)$
（確認を急ぐ ⇒ 急いで確認する）

特に近年は，(1a) のような語彙的言い換えを自動的に獲得する手法が精力的に研究されている [7, 8]．

しかしながら，そのような言い換え知識を適用する場面では，いくつかの課題が残されている．言い換え生成の側面から具体的に見てみると，次の 3 つの課題を克服する必要がある．

- 言い換え元の表現の曖昧性の解消
- 言語表現として適格な表現の生成
- 生成した表現が言い換え元の表現と同じ意味を持つ（意味的に等価である）ことの保証

言い換え元の表現が曖昧であっても，その曖昧性を保ったまま言い換えることができれば問題にはならないし，意味的等価性の検証に含めて処理することも考えられる．したがって，実質的には後者 2 つが課題となる．

これらの課題に対して，藤田ら [1] は「ある言語的に適格な表現 s に対して別の表現 t が言い換えである」ことを次の 3 つの条件で定義し，それらの充足度（**言い換えらしさ**）を推定する手法を提案している．

- t が言語的に適格である
- s が表すことが真ならば常に t が表すことも真である
- s を t に置換可能な文脈が存在する

提案されている手法は，2 つ目，3 つ目に関しては，共起表現の分布類似度によって細かい粒度でとらえようとしている．しかしながら，1 つ目の言語的適格性に関しては，その表現がウェブ上で 1 度以上用いられているかどうかという粗い近似に留まっている．

そこで本稿では，上の 3 つの条件で言い換えであることを定義するスタンスを踏襲しつつ，それらを統合的な確率モデルで表現する．なお，本稿では，我々のこれまでの研究 [4, 1] と同様，例 (2) に示すような言い換えを研究対象とする．

(2) a. 損害賠償を求める ⇒ 損害を賠償させる
b. 暖かい部屋だ ⇒ 部屋が暖まっている
c. デザインが魅力的だ ⇒ 魅力あるデザインだ

言い換え前の表現を構成する内容語（およびその派生形）の構造を変えることで実現できるこれらの言い換えを，我々は形態・構文的言い換えと呼んでいる．

2 言い換えらしさの定式化

我々は，ある述語句 s とその言い換え候補 t の言い換えらしさ $Par(s \Rightarrow t)$ を，次に示すように，s を条件とする t の生起確率として定式化する．

$$Par(s \Rightarrow t) \stackrel{\text{def}}{=} P(t|s).$$

ここで，s と t はパラディグマティックな関係にある（言い換えである）と仮定しているので，何らかの共起表現集合 F を考えると，上式は次のように変形できる．

$$\begin{aligned} P(t|s) &= \sum_{f \in F} P(t|f)P(f|s) \\ &= P(t) \sum_{f \in F} \frac{P(f|t)P(f|s)}{P(f)}. \end{aligned}$$

上式の第 1 因子 $P(t)$ は，言い換え候補表現 t の言語的適格性に対応する．一方第 2 因子は，s と t の共起表現の重複の度合いを表しており，共起表現の定義によって，表現間の意味的等価性あるいは置換可能性を表しうる．この節では以下，各因子について詳述する．

2.1 言語的適格性

言い換え候補表現 t の言語的適格性を表す $P(t)$ は，t の文節係り受け構造における各文節が $(N-1)$ 段までの係り先文節にのみ依存して生成されていると仮定する，構造を考慮した N-gram 言語モデルとする．今回は，各係り受けの独立性を仮定して（$N=2$），$P(t)$ を次式で与える．

$$P(t) = \left[\prod_{b \in T(t)} P_d\big(b|m(b)\big) \right]^{1/|T(t)|}.$$

*A Probabilistic Model for Generating Moorpho-syntactic Paraphrases.
Shuhei Kato†, Atsushi Fujita†, Satoshi Sato†
†Graduate School of Engineering, Nagoya University

表 1: 機能語部の抽象化の例
C=内容語．":" は形態素境界を表す．

係り元文節のタイプ	係り先文節のタイプ	削除する機能語（下線部）
名詞	動詞	C:が/C:ている:た 煙:が/あがる:ている:た （⇒ 煙:が/あがる）
名詞句（動）	形容詞	C:られる:た:の:が/C:た 認める:られる:た:の:が/うれしい:た （⇒ 認める:の:が/うれしい）

ここで，$T(t)$ は t に含まれる文節の集合，$1/|T(t)|$ は文節数による正規化項である．$P_d\big(b|m(b)\big)$ はある文節 b がすでに生成済みの文節 $m(b)$ に係る確率（以下，文節係り受け確率）を表し，次式で与える．

$$P_d(b_i|b_j) \quad = \quad \lambda_1 P_{d_1}(b_i|b_j) + \lambda_2 P_{d_2}(b_i|b_j).$$

ここでは，2 種類の文節係り受け確率を重み λ_1，λ_2 で混合している．P_{d_1} は文節の文字列そのものの係り受け確率であり，P_{d_2} は係り受け文節対のタイプ（機能語部に応じて判別する）に応じて係り元・係り先文節の機能語部を表 1 のように抽象化した場合の係り受け確率である．各々次式で与える．

$$P_{d_1}(b_i|b_j) \quad = \quad \frac{freq(b_i, b_j)}{freq(*, b_j)},$$

$$P_{d_2}(b_i|b_j) \quad = \quad \frac{abstfreq\big(b_i, b_j, type(b_i, b_j)\big)}{abstfreq\big(*, b_j, type(b_i, b_j)\big)}.$$

ここで，$freq(b_i, b_j)$ は所与のコーパス中で 2 つの文節 b_i と b_j が係り受け関係にある頻度，$abstfreq(b_i, b_j)$ は係り受け文節対のタイプに従って各文節の機能語部を抽象化した場合の係り受けの頻度，"$*$" は該当するあらゆる文節を表す．P_{d_2} を用いることで，文節の文字列そのものを用いる際のデータスパースネスの問題の軽減を図っている．文節係り受け確率の具体的な推定方法については 3.1 節で述べる．

2.2 意味的等価性および置換可能性

提案モデルの第 2 因子は，2 つの表現 s と t の意味的等価性（あるいは置換可能性）を，分布仮説に基づいて定量化したものである．すなわち，s と t と共起する表現の分布が似ているほど，両表現は意味的に類似している（あるいは置換可能である）とみなす．共起表現として文献 [1] で用いられた次の 2 種類を考える．

BOW: 表現と同じ文中の内容語．2 つの表現でこれらの分布が似ているほど意味的に類似していると仮定．

MOD: 表現の修飾・被修飾表現．2 つの表現でこれらの分布が似ているほど置換しやすいと仮定．

3 モデルの実装

3.1 文節係り受け確率

文節係り受け確率は，毎日新聞コーパス（1991-2005 年版）を用いて，次の手順で推定した．

表 2: 内容語の抽出パターン
正規表現で記述．ただし ":" は形態素境界を表す．

内容語の品詞	IPA 品詞体系における品詞
名詞	名詞–[^非自立]:名詞–接尾 {0,2} 形容詞–自立:さ
形容動詞	名詞–形容動詞語幹
動詞	動詞–自立 名詞–[^非自立]:名詞–接尾 {0,2}:(する | できる) 副詞–[^助詞類接続]:動詞–接尾 {0,2}:(する | できる) 形容詞–自立:がる
形容詞	形容詞–自立
副詞	副詞–[^助詞類接続]

1. 各文を MeCab[1] および CaboCha[2] によって解析し，依存構造を得る．
2. 各文節中の形態素列を表 2 に示したパターンに照らし，内容語（複数の場合もある）を同定する．パターンに合致しなかった部分を機能語部とみなす．
3. 次の 2 種類の表現対を係り受け文節対 $\langle b_i, b_j \rangle$ とみなし，頻度を $freq(b_i, b_j)$ とする．
   - 係り関係にある文節対から，各文節の最後尾の内容語以降の文字列を取り出し対とする．
     e.g. 最高記録を更新した ⇒ ⟨ 記録を, 更新した ⟩
   - 複数の内容語を含む文節中の，連続する内容語を取り出し対とする．ただし，末尾の内容語には機能語部を連結する．
     e.g. 株主総会決議を ⇒ ⟨ 株主, 総会 ⟩, ⟨ 総会, 決議を ⟩
4. 上述の係り受け文節対 $\langle b_i, b_j \rangle$ の各文節のタイプに応じて表 1 のような機能語部の抽象化を施し，これの頻度を $abstfreq\big(b_i, b_j, type(b_i, b_j)\big)$ とする．

3.2 述語句と共起する表現の生起確率

我々が対象とする述語句は，一般に語単体よりも生起しにくい．したがって，規模の小さなコーパスから十分な用例を得られるとは期待できない．そこで，2 つの述語句 s と t の意味的等価性・置換可能性の計算に用いる共起表現の確率分布 $P(f|s)$，$P(f|t)$ は，文献 [1] と同様，検索エンジンを用いて次の手順で推定する．

1. s および t そのものをクエリとして，Yahoo! JAPAN Web-search API[3] を用いてスニペットを取得する．
2. スニペットを形態素解析して BOW，係り受け解析して MOD に該当する表現をそれぞれ抽出する．これには文献 [1] で用いられたツールを援用したため，ChaSen[4] と CaboCha が用いられている．
3. 各表現の頻度から条件付き確率を算出する．

$P(f)$ の推定には，河原らがウェブから収集した 5 億文からなるコーパス [5] と毎日新聞の 2 種類を独立に用い，2 種類の確率モデルを構築した．

[1] http://mecab.sourceforge.jp/, ver. 0.96
[2] http://chasen.org/~taku/software/cabocha/, ver. 0.53
[3] http://developer.yahoo.co.jp/
[4] http://chasen.naist.jp/hiki/ChaSen/, ver.2.3.3

4　言い換え生成システム

我々はこれまで，統語構造変換パターンを用いて述語句の語彙・構文的言い換えを生成してきた [4, 1]．しかし，このアプローチでは，述語句の文法形式ごとに，パターンを記述する必要があり，知識の開発・メンテナンスに人的コストを要するという問題がある．また，実際の言い換え事例から一般化してパターンを記述している限りは，カバレージにも問題がある．

そこで，3 節で推定した確率モデルを用いて，入力された述語句 s を構成する内容語（およびその派生形）から統計的に生成可能なあらゆる言い換え候補 t を生成し，言い換えらしさのスコア $Par(s \Rightarrow t)$ とともに出力するシステムを試作した．言い換え候補の生成手順は次の 5 ステップからなる．

1. 内容語の抽出: 3.1 節で述べた文節係り受けモデルの作成ステップ 1，2 と同様にして，入力された述語句から内容語を抽出する．

2. 文節の生成: 抽出した内容語の各々に対して，派生語辞書 [3] を参照して派生語を列挙する．そして，元の語および派生形の各々に対して，機能語を付加して文節を生成する．ただし，付加する機能語の種類は，内容語の品詞ごとに表 3 に示すものに限定した．生成した表現に応じて表 3 のように文節のタイプを定める．これは，係り受け確率を計算する際の機能語部の抽象化（表 1）に利用する．

3. 文節係り受け構造の生成: 生成した文節を組み合わせることで生成可能なあらゆる係り受け構造を生成する．ここでは，入力中のすべての内容語（およびその派生語）を必ず 1 度ずつ用いるという制約を満たす係り受け構造を網羅的に生成する．ただし，文節係り受け確率が 0 となるような文節対を含む係り受け構造は生成しない．

4. 表層生成: 言い換え候補表現の係り受け構造を文字列化する．各活用語の活用形を修正し．姉妹文節の語順は後から生成された文節を前にする．

5. 言い換えらしさの計算: 各言い換え候補表現 t に対して，検索エンジンの API を通じてスニペットを収集し，言い換えらしさのスコア $Par(s \Rightarrow t)$ を計算する．

5　言い換え生成実験

我々は，2 節で述べたように，ある述語句 s とその言い換え候補 t の言い換えらしさ $Par(s \Rightarrow t)$ を，s を条件とする t の生起確率 $P(t|s)$ として定式化した．これに照らし，ある表現に対する複数の言い換えをどれくらい適切にランキングできるかを評価するための言い換え生成実験を行った．

表 3: 生成する機能語の範囲
正規表現で記述．ただし ":" は形態素境界を表す．
また，C=当該内容語，が=(が｜を｜に｜で)，こと=(こと｜の)，た=(た｜だ)，させる=(させる｜せる)，られる=(られる｜れる)，ている=(ている｜でいる)

内容語の品詞	文節のタイプ	扱う機能語の範囲
名詞，形容動詞	述語（名） 名詞	C:だ:た* C:が
副詞	述語（副） 副詞	C:だ:た* C:が*
形容詞	形容詞 名詞句（形）	C:た* C:た*:こと:が
動詞	動詞 名詞句（動）	C:させる*:られる*:ている*:た* C:させる*:られる*:ている*:た*:こと:が

5.1　比較するモデルの一覧

提案モデルには，$P(f)$ の推定方法が異なる 2 種類がある．以下では，ウェブコーパスを用いたモデルを *WebCP*，毎日新聞を用いたモデルを *Mainichi* と記す．また比較対象として，文献 [1] で用いられた Lin の類似度 [7]（以下，*Lin*）と α-skew divergence [6]（以下，*skew*）の 2 種類を用いた．

提案モデル，比較モデルのそれぞれについて，言い換えらしさの推定に BOW，MOD という共起語集合を独立に用いるモデルとそれらの推定値の調和平均（HAR）の 3 種類を考えた．さらに，生成した表現のヒット数を言い換えらしさとするベースライン（以下，*HITS*）の，合計 13 種類のモデルを比較した．

5.2　評価用データセットの作成

文献 [1] で言い換え生成実験に用いられた 6,190 個の述語句[5]に対して言い換え候補表現を生成した．候補生成時の条件，およびパラメタ設定は次の通りである．

- 述語句のみを出力する
  - 末尾がタ形の候補を出力しない
  - 末尾の接尾辞が変化するだけの候補を出力しない
  - 入力と姉妹順序が異なるだけの候補を出力しない
  - 上の 2 つの組み合わせも出力しない
- 文節係り受けモデルにおける重み λ_1,λ_2 は共に 0.5
- スニペットは最大 1000 件を取得

結果，5,292 個の表現に対する 280,712 件の言い換え候補（平均 53 件）を得た．入力した述語句のうち 3,851 件に対しては，13 種類のモデルすべてが 1 件以上の言い換え候補を出力した．この入力述語句の中からランダムに 200 件を選択し，各モデルが 1 位とした言い換え候補 200 件ずつ（異なりで 434 件）を評価用のデータセットとした．

[5] 6 種類の文法形式の述語句からなる．例を示す．N:C:V…確認を急ぐ，N_1:N_2:C:V…出火原因を調べる，N:C:V_1:V_2…統計を取り始める，N:C:Adv:V…検討をさらに進める，Adj:N:C:V…高い評価を受ける，N:C:Adj…のどが痛い

表 4: システムの出力例と判定結果

言い換え元の述語句 s ⇒ 言い換え候補表現 t	2 名の判定結果
法案に反対する ⇒ 法案に反対だ	言い換えである
数が少ない ⇒ 数が少なめだ	言い換えである
強い反発が出る ⇒ 反発が強く出る	言い換えである
意識が薄い ⇒ 意識が薄れている	言い換えでない
捜査本部を置く ⇒ 捜査本部が置かれている	言い換えでない
指導要領に基づく ⇒ 要領に基づき指導する	言い換えでない

5.3 評価結果と考察

434 件の言い換え候補対が正しい言い換え対であるかどうかを，2 名の被験者が独立に判定した．ただし，入力の述語句 s は機械的に抽出したものであるため，不適格な表現を含んでいる可能性がある．そこで，1 節の 3 つの条件に加えて，s が言語的に適格であるかどうかも判定した．2 名の判定が一致した例を表 4 に示す．

2 名ともが言い換え前後の表現 s，t ともに適格であると判定したは 276 件 (64%) であった．これらの言い換え候補対に対する 1 節の 2 つ目，3 つ目の条件（言い換えか否か）の判定結果の一致率は 84%，κ 統計量は 0.65 であった．ただし，2 名の被験者のうち片方のみが『言い換えである』とした事例の数には，37 件，7 件と偏りがあった．これは，1 名が 3 つ目の条件を “あらゆる文脈で置換可能” と誤解して，より多くの言い換え候補を『言い換えでない』としたために生じていた．

被験者の 1 名以上が言い換えであると判定した候補を最終的に『言い換えである』としたところ，各モデルの精度は表 5 に示すようになった．今回比較したモデルの精度は最高でも 37%であり，文献 [1] で報告されている同種の実験の 62〜67%という結果に比べて大幅に低い．これは，生成する言い換えの種類に関する制約に起因している．文献 [4, 1] では，言い換え候補として例 (3) のような内容語の数が減る言い換えも生成していた．

(3)　窓ガラスを割る ⇒ ガラスを割る

このような言い換えでは各要素の名詞句が一般化されるだけなので，ほとんどの場合に言い換えであると判定されていた．しかし，今回提案した言い換え生成モデルでは，このような言い換えを生成しない制約（4 節のステップ 3）を設けてあった．このため，ヴォイス，アスペクト表現のような細かい意味の差を持つ表現を付加した言い換え候補が相対的に増加していた．しかし，これらはほとんどの場合正しい言い換えとはならないので，相乗的に精度を下げてしまったと考えられる．

最後に，1 位出力 200 件中の，t が適格だと判定された件数と割合を表 6 に示す．提案モデルは，統計的に優位ではないが，比較モデルよりも言語的に適格な表現を上位に出現させる能力が高かった．しかしながら，

表 5: 1 位出力 200 件中の正しい言い換え候補の数

	BOW	MOD	HAR
WebCP	58 (29%)	73 (37%)	73 (37%)
Mainichi	65 (33%)	66 (33%)	70 (35%)
Lin	64 (32%)	72 (36%)	73 (37%)
skew	69 (35%)	72 (36%)	72 (36%)
HITS	68 (34%)		

表 6: 1 位出力 200 件中の適格な表現の数

	BOW	MOD	HAR
WebCP	146 (73%)	145 (73%)	147 (74%)
Mainichi	143 (72%)	140 (70%)	143 (72%)
Lin	133 (67%)	120 (60%)	131 (66%)
skew	133 (67%)	129 (65%)	130 (65%)
HITS	133 (67%)		

言い換えらしさ全体の評価（表 5）ではこの強みが失われてしまっていることから，類似性のモデル化についても改善が必要であると考えられる．

6 おわりに

本稿では，言い換えらしさを確率的にモデル化し，述語句に対する言い換え生成実験によってモデルの妥当性を検証した．結果，提案モデルは比較モデルより，言語的に適格な表現を上位に出現させる能力が高いことが確認できたが，言い換えらしさの推定精度は全体的に低かった．今後は，$P(t)$ モデルのノイズ除去やスムージング，類似度のモデル化の改善に取り組む予定である．

本研究の一部は次の科研費の支援を受けている：科研費基盤研究 (A)「円滑な情報伝達を支援する言語規格と言語変換技術」（課題番号：16200009，代表：佐藤理史）および科研費若手研究 (B)「文法カテゴリ交替を裏付ける語彙特性の体系化と辞書記述」（課題番号：18700143，代表：藤田篤）

参 考 文 献

[1] 藤田篤, 佐藤理史. 述語句統語的異形間の言い換えらしさの計算手法. 情報処理学会研究報告, NL-182-4, pp. 23–30, 2007.

[2] 乾健太郎, 藤田篤. 言い換え技術に関する研究動向. 自然言語処理, Vol. 11, No. 5, pp. 151–198, 2004.

[3] 加藤直樹, 藤田篤, 佐藤理史. 語末の形態的特徴に基づく日本語派生語対の収集. 言語処理学会第 13 回年次大会発表論文集, pp. 352–355, 2007.

[4] 加藤修平, 藤田篤, 佐藤理史. 句を対象とした構成的な言い換えの生成. 言語処理学会第 13 回年次大会発表論文集, pp. 903–906, 2007.

[5] 河原大輔, 黒橋禎夫. 高性能計算環境を用いた Web からの大規模格フレーム構築. 情報処理学会研究報告, NL-171-12, pp. 67–73, 2006.

[6] L. Lee. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 25–32, 1999.

[7] D. Lin and P. Pantel. Discovery of inference rules for question answering. *Natural Language Engineering*, Vol. 7, No. 4, pp. 343–360, 2001.

[8] I. Szpektor, H. Tanev, I. Dagan, and B. Coppola. Scaling Web-based acquisition of entailment relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Lang uage Processing* (*EMNLP*), pp. 41–48, 2004.