動詞間の換言知識の自動獲得

仁井 康夫　　酒井 浩之　　吉田 辰巳　　増山 繁

豊橋技術科学大学知識情報工学系
〒441-8580 豊橋市天伯町雲雀ヶ丘1-1

E-mail : {nii,sakai,gaizi}@smlab.tutkie.tut.ac.jp, masuyama@tutkie.tut.ac.jp

1 はじめに

換言とは，同一の意味を保ったまま異なる表現に言い換えることである．我々は，会話や文章の記述に当り，この換言処理を日常的に行っている．そのため，換言は，人間と同様の高度な翻訳や要約を機械的に行うために，必要不可欠な要素技術の一つといえ，関連研究も多数存在する [1].

また，動詞は一文中にほぼ必ず含まれる品詞であり，文章の構成要素としての重要度は極めて高い．さらに，例1に示すように，共通の意味を含んだ異なる表現が多く存在し，文章の記述においても様々な使い分けが行われている．したがって，動詞に関する換言知識を取得することは，検索や要約，情報弱者のための難解語の言い換え等，様々な分野に応用することができる．

例1

話す，語る，会話する→ 言う

動詞の換言には，先行研究 [2][3] が存在する．しかし，いずれも，換言先動詞は，EDR日本語単語辞書や国語辞典の定義文，もしくは人手によって選択されている．したがって，これらの方法で得られた換言語は，網羅性，新語への対応，換言先動詞を追加する際の個人差による判断の揺らぎ等の問題があると考える．

そこで，本研究では，動詞が出現する文脈から得られる統計情報を基に，動詞間の換言可能性を機械的に定量化する手法を提案する．提案手法では，言い換えられる動詞の候補を，大量のコーパスから統計情報を用いて機械的に発見する．さらに，日本語特有のヒューリスティックスを用い，被換言動詞が含む漢字の情報より，統計情報のみでは省けない換言不可能語の削減や類似度の補正を図る．そのため，提案手法では，換言候補の発見におけるコスト低下，網羅性，新語への対応を期待できると考える．

2 統計情報による類似度の計算

本研究では，大量のコーパスを基に，各動詞に関する統計情報を素性とするベクトルを作成し，それらのベクトルを比較することにより動詞の換言可能性を定量化する．これは，似た状況で出現することが多い動詞間には，関連性があるだろうとの仮定に基づく．ただし，動詞に関する統計情報には様々なものがある．例えば，動詞と同一文中に出現する名詞の種類や，その意味素性等の出現回数などである．その中でも今回は，換言対象が動詞であることを考慮し，主として構文的にその動詞に係る文節内の情報を用いる．さらに，補助的な情報として，動詞が同時に取る表層格の組み合わせとその頻度を用いる．

2.1 動詞に係る名詞+表層格による類似度の計算方法

動詞 v に対応するベクトルの素性 i は，その動詞に係る名詞+表層格を単位として作成する．このとき，素性 i の重み $w_v(i)$ は，動詞とそれに係る名詞+表層格との係り受け回数により，以下の式で計算する．

$$w_v(i) = \frac{f(v,i)}{\log(f(v)+f(i))} \log \frac{V}{vf(i)} \qquad (1)$$

$f(v,i)$：コーパス中で名詞＋表層格の組 i が動詞 v に係る回数
$f(v)$：コーパス中での動詞 v の出現回数
$f(i)$：コーパス中での名詞＋表層格の組 i の出現回数
$vf(i)$：名詞＋表層格の組 i が係る動詞の種類の数
V：コーパス中に出現する動詞の種類の数

ベクトルAとBの類似度は，一般的に式 (2) に示す余弦を比較することによって行われる．

$$sim(A,B) = \frac{ab}{|a||b|} \qquad (2)$$

しかし，多くの換言可能な動詞間は，例2に示すようにその方向が一方向である場合が多い．

例2

○出土する → 見つかる
×見つかる → 出土する

しかしながら，動詞間の類似度をベクトルの余弦によって判定したのでは，その結果がいずれの方向の換言を可能としているのかは判断できない．

さらに，動詞が取る意味の範囲を考えると，図1に示すように，意味の狭い動詞がより広い意味を持つ動詞に含有されている場合が多く存在することが分かる．したがって，動詞間の換言可能性は，図1の「要請と求める」や「要請と頼む」の関係のように，換言先動詞

のベクトルが被換言動詞のベクトルをどの程度カバーしているかによって決定する方が有効であると考えられる．そこで，動詞 A と動詞 B の類似度は，以下の式を用いて計算する．

$$Sim_{nc}(A,B) = \sum_i \frac{a_i * f(a_i,B)}{|a|} \quad (3)$$

ただし，$f(a_i,B)$ は，ベクトル B の素性内に，ベクトル A の素性 a_i と同一の素性が存在した場合に 1 を，それ以外は 0 を表すバイナリ関数とする．なお，バイナリ値を利用した理由は，今回統計情報を取得した新聞記事コーパスの場合，係り受け関係にある名詞とその出現頻度を特徴とすると，必ずしも換言可能な動詞が被換言候補の特徴を十分に内包していないことが多数存在するからである．例えば，「飛行機が墜落した」という文はコーパス中に多数出現し，「飛行機が」は「墜落する」のベクトルを表す大きな特徴となっている．一方，「飛行機が落ちた」という文はコーパス中にほとんど出現しない．そのため，名詞とその出現頻度を特徴とした場合，「落ちる」は「墜落する」が持つ大きな特徴を共有していないこととなり，換言不可能と判断されてしまう．

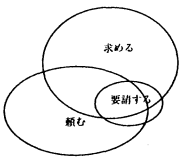

図 1: 動詞が持つ意味の範囲の例

2.2 動詞が取る表層格の組み合わせを用いた類似度の計算方法

提案手法では，動詞に係る名詞+表層格以外に，補助情報として，各動詞が取る表層格の組み合わせの頻度を用いる．そのため，動詞 v に同時に係る，文節内に含まれる表層格の組み合わせとその頻度をコーパスより調べ，各組み合わせを素性，その頻度を重みとして持つベクトルを作成する．なお，この方法による動詞 A と動詞 B の類似度 $Sim_c(A,B)$ の計算は，上記の方法と同様に式 (3) によって行う．

2.3 動詞間の類似度の判定

上記に示した 2 つの類似度を基に，式 (4) によって動詞間の類似度 $Sim(A,B)$ を定量化する．

$$Sim(A,B) = Sim_{nc}(A,B) * Sim_c(A,B) \quad (4)$$

3 換言候補の削減

換言候補をあらかじめ削減することにより，2 節で示した手法の精度の向上を図る．

3.1 類似文からの換言候補の取得

この手法は，類似する文中で出現する動詞は，類似している場合が多いとの仮説に基づく．すなわち，被換言動詞を述部において使用する文に対して，同一コーパス中より類似文を検索し，それらの類似文中で使用されている語を換言候補とする．以下にその手順を示す．

Step 1 被換言動詞 x を述部に持つ文集合 S_x をコーパスより集め，S_x 内で使用されている名詞集合 $T(x)$ を得る．

Step 2 名詞 $t_{S_x}(i) \in T(x)$ に対して，以下の式により順位付けを行い，n 以下の名詞を除く．

$$w_n(t_{S_x}(i)) = tf(t_{S_x}(i)) \log \frac{N}{vf(t_{S_x}(i))} \quad (5)$$

ただし，$tf(t_{S_x}(i))$ は名詞 $t_{S_x}(i)$ の文集合 S_x での出現回数，N はコーパス内の動詞の種類，$vf(t_{S_x}(i))$ はコーパス内で名詞 $t_{S_x}(i)$ が係る動詞の種類を表す．

Step 3 動詞 x を含む文とコーパス中の各文 s_y との類似度を，先に求めた名詞 $t_{S_x}(i)$ の重み $w(t_{S_x}(i))$ と文 s_y 中に出現する名詞 $t_{s_y}(i)$ との一致によって，以下の式で求める．そして，類似度 $sim(T(x),s_y)$ の上位 m 位までの文集合 S_{sim} を得る．

$$sim(T(x),s_y) = \frac{\sum_{t_{S_x}(i)=t_{s_y}(j)} w_n(t_{S_x}(i))}{\sum_i w(t_{S_x}(i))} \quad (6)$$

Step 4 文 $s_y \in S_{sim}$ の述部で使用されている動詞 y を取得し，以下の式によって，求められた重み $w_v(y)$ の上位 l までを換言候補とする．

$$w_v(y) = \sum_i 0.8^{1/sim(T(x),s_y)} * f(s_i,y) \quad (7)$$

ただし，$f(s_i,y)$ は文 s_i 内に述部として動詞 y が出現していれば 1，それ以外は 0 を表す．

3.2 動詞の意味の広さによる削減

先に述べたように動詞の換言例を考えると，狭い意味を持つ動詞から広い意味を持つ動詞への換言が可能であるのに対し，多義性を解消しない限り，その反対方向の換言は困難である．提案手法では，動詞の多義性の解消方法については考慮していないため，被換言動詞が持つ意味よりも広い意味を持つ動詞を換言候補よりあらかじめ省くことによって，多義性の影響をできる限り排除する．

動詞が持つ意味の広さを判定するためには，その動詞が持つ漢字に着目すればよい．日本語の動詞のほとんどには，漢字が含まれており，含まれている漢字の種類によって，以下の 3 つのパターンに分類できる．

1. 漢字+かな
   例：進む，出る，話す，言う，求める等
2. 動作を表す漢字+動作を表す漢字
   例：進出する，抽出する，解決する，取り扱う等
3. 動作を表す漢字+その目的語として使われる漢字
   例：見物する，撮影する等

このとき，複数の動作を持つ動詞は，単一の動作によって表される動詞よりも意味が狭いと言える．また，目的語を含む動詞は，その動作がさらに限定されるため，複数の動作を含む動詞よりも，さらに意味が限定すると考えられる．よって，各分類に含まれる動詞の広さは，1. > 2. > 3. の関係を持つと思われる．そこで，提案手法では，被換言動詞の種類によって，3. → 3.,2.,1.，2. → 2.,1., 1. → 1. のみを換言候補とする．

なお，動詞 v の分類は以下のようにして決定する．

if 動詞が漢字 1 文字+かな
then 動詞 v の分類は 1. 終了.
for 動詞 v の i 番目の漢字 $c(v,i)$ について
  if v_i が形態素解析済みコーパス内で" 漢字 $c(v,i)$+かな" の形で動詞として使われている
  then 動詞 v の分類は 2. 終了.
動詞 v の分類は 3. 終了.

4 ヒューリスティックスによる類似度の補正

動作を表す漢字を含む動詞の場合，例 3 に示すように，その動詞が含む動作を表す漢字を用いて，「漢字+かな」で表される動詞 (以下和語動詞と呼ぶ) に換言可能な場合が多い．また，このことは，和語動詞への換言のみではなく，一般の動詞への換言の場合についてもいえる．

例 3

表現する → 表す

そこで，被換言動詞と共通の漢字を含む動詞の類似度を上げることにより，統計情報による類似度のみでは取得できなかった換言候補の取得が期待できる．ただし，この方法では，被換言動詞が含む動詞のみしか類似度を上げることができず，出力される換言候補の再現性を逆に低下させる恐れがある．そこで，このヒューリスティックスを，2. で求めた換言候補の順位と類似度に基づいて適用することにより，被換言動詞に含まれる以外の漢字を含む換言先動詞についても，類似度の補正を試みる．以下にその手順を示す．

Step 1 2. で示した手法によって求めた類似度により，全換言候補の順位づけを行う．ただし，一位には被換言語自身が来るようにする．また，動詞 v の被換言動詞との類似度を $score(v)$ とする．
Step 2 上位 n 位までの換言候補に対し，以下の処理を繰り返す．
Step 2-1 上位 i 位の換言候補 v_i と同一の漢字を持つ換言候補 v_j のスコア $score(v_j)$ に対し，$score(v_i)/(i+1)$ を加える
Step 3 スコアによって換言候補の順位を入れ換える．

5 実験と評価

5.1 実験

以上の手法を実装し実験を行った．コーパスには毎日新聞 98 年の約 120000 記事を，形態素解析器には JUMAN を，構文解析器には KNP を用いた．また，被換言動詞は，同コーパスより，出現頻度が 10 回以上のサ変名詞を，異なり数が 100 になるまでランダムに選択することによって決定した．さらに，換言候補は，同コーパス中に出現する動詞のうち，出現回数が 5 回以上の約 15000 語を用いた．

5.2 評価

2 節と 3 節で説明した手法によって，被換言動詞と換言候補間の類似度を求め，その値が閾値以上の上位 10 位までを換言先動詞として出力した場合の実験結果を図 2 に示す．

図 2 より分かるように，提案手法では 100 個のサ変名詞に対し，閾値 0.3 において，133 個の換言候補を出力し，精度は 35.6%であった．なお，提案手法によって得られた換言例を表 1 に示す．

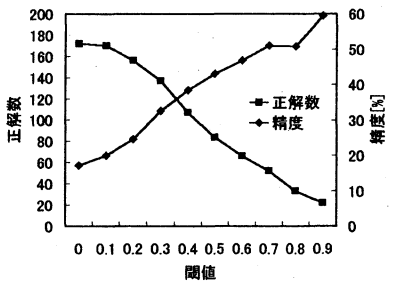

図 2: 閾値と換言結果の精度の関係

一方，失敗の例としては，例 4 に示すような，換言はできない関連語，反対語，動詞と他動詞の違い等が見られた．

例 4

×逮捕する → 起訴する

表 1: 獲得できた換言例

被換言動詞	換言動詞
急騰する	値上がる，上昇する，急伸する
出場する	出る，参加する，出席する
可決する	採択する，了承する，決まる
更新する	上回る，塗り替える

×急騰する → 値下する
×整備する → 整う

また，各補助手法の効果を示すために，各補助手法を除いた場合，正解数 100 を得ることができる精度の比較を表 2 に示す．

表 2: 正解数 100 における各補助情報と精度の関係

使用しない補助情報	精度 [%]
なし	40.7
表層格の組み合わせ	35.5
換言候補の削除	27.3
類似度の補正	26.8

表層格の組み合わせを省くことによって 5%，換言候補の削減を省くことによって 13%，さらに類似度の補正を省くことによって 14%の低下がそれぞれ見られた．これより，動詞間の換言候補を決定する際に，各補助手法は有用であったといえる．

6 今後の課題

提案手法では，構文解析器によって得られるの知識以外，文法に関する知識を一切利用していない．しかし，動詞に関する文法知識を利用すると，あらかじめ換言候補を削減することができる．

利用する動詞の分類としては，例えば，金田ら [4] による 4 分類や，日本語語彙大系による用言意味分類 [5] 等を考えている．これらを基に，動詞の分類を教師あり学習させることにより，統計情報以外による動詞の分類が可能であると考える．

また，新聞記事では使用状況が限定され，十分な統計情報を取得できない動詞間の換言知識については，ニュース原稿や小説等の他分野のコーパスからの統計情報との併用が有効であると考える．これは，ニュース原稿や小説等では，同一の言い回しを異なる表現で言い換えられている例が多数存在するからである．

さらに，現在では，その動詞が持つ意味の範囲が狭いものから広いものへの換言のみを考えているが，その反対方向への換言への換言知識の獲得も重要であると考える．このとき，問題になることは動詞の多義性の解消である．しかし，統計情報を用いた多義性の解消手法はすでに多く提案されており [6]，これらを本手法に組み込むことにより，反対方向への換言知識の獲得は可能であると考える．

7 まとめ

本研究では，動詞に係る名詞と表層格に関する統計情報を基に作成したベクトルを利用し，動詞間の換言知識の自動獲得を試みた．その結果，動詞に含まれる漢字を利用したヒューリスティックスと，換言候補を被換言動詞を含む文の類似文より選択することにより，100 個のサ変名詞に対し 133 個の換言動詞と，精度 35.6%の性能を得た．

今後の課題としては，同分野のコーパスでは使用状況が限定され，換言候補を抽出できない動詞に対する他分野のコーパスを適用した換言候補の取得方法，動詞の分類による更なる換言候補の削減の構築，多義性を用いた双方向の換言知識の獲得，さらに，ベクトルの重みの改良による精度の向上が上げられる．

謝辞

本研究は文部科学省２１世紀 COE プログラム「インテリジェント　ヒューマンセンシング」，及び，日本学術振興会科学研究費基盤研究 (C)(2)13680444 の援助により行われた．

また，言語データとして，毎日新聞 CD-ROM 版の使用を許可して頂いた毎日新聞社に深謝する．

参考文献

[1] 乾健太郎：言語表現を言い換える技術，言語処理学会年次大会第 8 回年次大会チュートリアル資料，pp. 1-21(2002)

[2] 近藤恵子，佐藤理史，奥村学,「サ変名詞+する」から動詞相当句への言い換え，情報処理学会論文誌，Vol.40,No.11, pp.4064-4074, 1999.

[3] 鍛冶伸裕，河原大輔，黒橋禎夫，佐藤理史，国語辞典とコーパスを用いた用言の言い換え規則の学習，言語処理学会第 8 回年次大会発表論文集，pp.331-334,2002.

[4] 金田一春，林大，柴田武，日本語百科大辞典，大修館書店, 1989.

[5] 池田悟 他，日本語語彙体系，岩波書店, 1997.

[6] Christopher D. Manning, Hinrich Schütze : Foundations of Statistical Natural Language Processing, The MIT Press Cambridge, Massachusetts London, England, pp.229-264, 1999