# 理解補助を目指した動詞句の換言

大田 浩志 , 山本 和英
長岡技術科学大学 電気系
E-mail:{ota,ykaz}@nlp.nagaokaut.ac.jp

## 1 はじめに

換言技術の用途のひとつに人間の理解補助がある。例えば、言語初学習者を考えると語彙不足が理解を困難にしていることは明らかである。テキストによる情報を円滑に伝達する際には、読み手の言語能力にあわせた読みやすい平易な文にすることが必要である [1]。

例 1) 彼は煙草を吸おうと マッチを擦る。
彼は煙草を吸おうと 火をつける。

例 1 に示す 2 文は概ね同義である。これは「マッチを擦る」と「火をつける」という同義ではない動詞句が換言表現となることを示している。語彙の異なりは文の読みやすさ平易さに影響を与える。例にあげた動詞句対のように原文の意味を概ね保持したままでの換言が可能な表現対を収集することは、読みやすさ平易さの異なる換言表現の獲得に繋がる。これは理解補助のための換言に有効であると考える。

「やさしい日本語」[3] では災害時において外国人にもわかりやすく日本語で情報を伝えるための方策を提案している。その中では「暖かくする」を「服をたくさん着る」とする換言表現を提案している。これらの表現はそれぞれの意味する情報が一致しないため、換言可能な状況は制限される。しかしこの換言には概ね同義で原文より円滑な情報伝達を実現できる可能性がある。このような換言表現を機械的に抽出することが本稿の目的である。

「マッチを擦る」と「火をつける」の動詞句間には「マッチを擦って火をつける」「マッチを擦り火をつける」のように係り受け関係が成立する。本稿では、係り受け関係にある動詞句は換言表現となる可能性があると考え、ある条件下で換言可能な表現を抽出する手法を提案する。

仁井ら [4] は動詞に係る格要素に関する統計情報を用いて、動詞間の換言を行っている。長谷川ら [5] は大規模コーパスを使い、固有表現との位置関係から換言表現を抽出している。本稿では、述語に係る格要素の統計情報及び換言表現周辺の単語を一致度を指標として用いて、動詞句の換言を行った。

## 2 提案手法

### 2.1 手法概要

本手法の処理は前処理部と選出部の 2 つに分けられる。前処理部では、 図 1 に示す 4 つ組を収集する。次に収集した 4 つ組を整形し、係り受け関係にある動詞句対として保持する。係り先動詞句は係り元動詞句の換言候補として扱う。4 つ組、整形方法、動詞句対については 2.2 節で述べる。

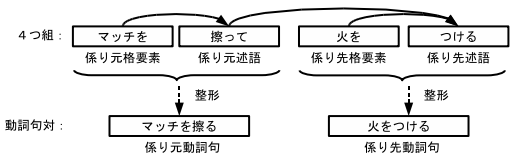


図 1: 4 つ組と動詞句対例

選出部では換言候補中から複数の尺度を用いて換言となる動詞句の選出を行う。換言を選出する尺度として名詞-名詞頻度スコア、動詞-名詞頻度スコア及び周辺単語一致度スコアを用いた。これらのスコアについては 2.4.1 節及び 2.4.2 節で述べる。

### 2.2 前処理部

構文解析[1] を行ったコーパスを用いて動詞句対を収集する。本稿で使用する 4 つ組の定義を行う。動詞を含む文節を述語、名詞を含む文節を格要素とする。図 1 に示す「擦って」と「つける」のように互いに係り受け関係にある述語をそれぞれ係り元述語、係り先述語とする。「マッチを」のように係り元述語に係る格要素を係り元格要素、「火を」のように係り先述語に係る格要素を係り先格要素とする。[係り元格要素：係り元述語：係り先格要素：係り先述語] の組を 4 つ組として収集する。

格要素の名詞[2]は「名詞-一般」、「名詞-サ変接続」及びそれらによる複合名詞とする。
また述語には次に示すような動詞の整形を施す。

- 未然形以外の動詞は原形にする。
- 動詞が 2 つ以上連続する場合は、1 つの動詞として扱う。
  例 2) 「食べ残し」→「食べ残す」
- 未然形の動詞は表層形を使用し、それに続く助動詞は原形にする。未然形の動詞とそれに続く助動詞はまとめて 1 つの動詞として扱う。
  例 3) 「食べなくなる」→「食べないなる」
- 読点が続く未然形動詞は原形を使用する。
  例 4) 「属し、」→「属する、」
- 動詞直前の形容詞、副詞、名詞及び名詞に連続する助詞は表層形を使用する。それ以外の形態素は削除する。
  例 5) 「『幅広くし」→「幅広くする」

係り元述語は 2 つ以上の動詞を含まないものとする。また、整形を行った動詞に形態素が続く場合は助詞もしくは読点のものとする。該当しない場合は収集対象としない。係り先述語は 2 つ以上の動詞がある場合、1 つ目の動詞のみを整形対象とする。述語では、整形を行った動詞以降の形態素は削除する。

整形した 4 つ組において、係り元格要素及び係り元述語から出来る動詞句を係り元動詞句とする。係り先格要素及び係り先述語から出来る動詞句を係り先動詞句とする。これらを動詞句対として保持する。係り元述語から削除した助詞は動詞句間を繋ぐ意味があり、換言となる動詞句対において特徴があると考える。そこで、動詞句間を繋ぐ助詞として保持した。

### 2.3 換言元動詞句

換言元動詞句を抽出する。まず係り受け関係にある格要素と述語を抽出する。格要素と述語は連続して出現するものに限る。本稿ではヲ格の格要素を持つ換言元動詞句に限

[1]本稿では係り受け解析に CaboCha(1) を用いる。
[2]本稿で扱う品詞情報には ChaSen(2) の品詞体系を用いる。

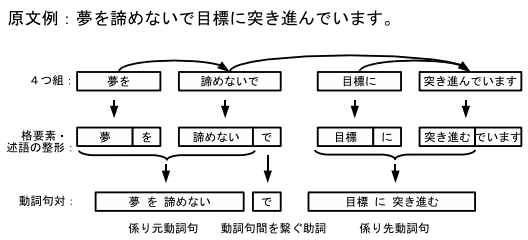

図 2: 4 つ組から動詞句対への整形

る。この理由は換言表現となる動詞句対を繋ぐ助詞は特徴があると考え、その助詞は換言元動詞句の格助詞による影響を受けると考えるためである。格要素、述語の品詞の条件はそれぞれ 2.2 節の格要素及び係り元述語と同様とする。また、整形方法についても同様である。整形した格要素及び述語から出来る動詞句を換言元動詞句とする。

### 2.4 選出部

選出部では与えた換言元動詞句に対する換言動詞句を出力する。換言元動詞句が係り元動詞句となっている動詞句対を全て抽出する。抽出した動詞句対の係り先動詞句を換言元動詞句の換言候補として扱う。各言い換え候補に 3 つの異なるスコアを付与する。各スコアに閾値を定めてフィルタリングし候補中から換言動詞句の選出を行う。

#### 2.4.1 名詞-名詞頻度スコア・動詞-動詞頻度スコア

名詞-名詞頻度スコアでは、まず換言元動詞句 $S$ と同一の格要素 $N$ である係り元動詞句の動詞句対を全て抽出する。表 1 は換言元動詞句を「意見を聞かない」とした際に抽出される動詞句対の例である。格要素が「意見を」である係り先動詞句の動詞句対が抽出されている。続いて係り先動詞句で用いられている名詞 $n$ に対して抽出した全動詞句対を対象に頻度を求める。本稿ではこれを名詞-名詞頻度 $f_N(n)$ とする。表 1 の場合係り先動詞句で用いられている各名詞の $f_N(n)$ は「考え」が 2、「独断」が 2、「病室」が 1 である。次にこの $f_N(n)$ を用いて換言候補へのスコア付けを行う。

表 1: $f_N(n)$ 算出のための動詞句対 $S$:意見を聞かない

| 係り元動詞句 | | 係り先動詞句 |
|---|---|---|
| 意見を聞かない | ⇔ | **考え**を押しつける |
| 意見を述べる | ⇔ | **考え**を伝える |
| 意見を聞く | ⇔ | **独断**で決める |
| 意見を聞かない | ⇔ | **独断**で決定する |
| 意見を聞かない | ⇔ | **病室**へと歩き始める |

換言候補 $S'$ に含まれる名詞 $n'$ の名詞-名詞頻度 $f_N(n')$ を名詞スコア $p(S')$ として $S'$ に付与する。観察結果より、実際に換言可能な候補は他の候補に比べ $p(S')$ が大きくなりやすい傾向がある。そこで他の換言候補と相対して $p(S')$ が大きい候補ほど正解らしいとしスコアが高くなるように再度スコア付けを行う。

名詞スコア $p(S')$ の昇順で $S'$ に全換言候補 $S'_{all}$ 内で順位 $r(S')$ をつける。$p(S')$ が等しい場合は低い順位で同位とする。例えば最も高い $p(S')$ を持つ 2 つの $S'$ は共に $r(S')$ が 2 となる。各換言候補に順位を換言候補数 $|S'_{all}|$ で割った値を付与する。これを各換言候補の名詞-名詞頻度スコア $NN(S')$ とする。$NN(S')$ は他の全ての換言候補との間で相対的に決めているスコアである。換言候補数が少ない場合の順位及び頻度によるスコアは信頼性が低いと考える。そこで換言候補の $p(S')$ の総計 $P_{S'_{all}}$(式 (1)) が 10 以下であるときその数に応じて減点する (式 (2))。ただしここでの減点方法は試行により決定している。減点は $P_{S'_{all}}$ が 1 のとき $NN(S')$ の最大値が 1/2 となるように設定した。名詞-名詞頻度スコアのとる値は $0 < NN(S') \leq 1$ である。

$$P_{S'_{all}} = \sum_{S' \in S'_{all}} p(S') \tag{1}$$

$$NN(S') = \begin{cases} \frac{r(S')}{|S'_{all}|} \cdot \frac{1}{100}(45 + 5 \cdot P_{S'_{all}}) & if\ P_{S'_{all}} \leq 10 \\ \frac{r(S')}{|S'_{all}|} & otherwise \end{cases} \tag{2}$$

表 2 は「意見を聞かない」に対する各 $S'$ に付与する $p(S')$,$r(S')$,$NN(S')$ の例である。「病室へと歩き始める」の名詞である「病室」は $p(S')$ が 1 で $r(S')$ は 1 位。「考えを押しつける」と「独断で決定する」の名詞である「考え」、「独断」の $p(S')$ は同じでどちらも $r(S')$ は 3 位とする。各言い換え候補の順位 $r(S')$ を換言候補数 3 で割り $NN(S')$ を算出している。$p(S')$ の総計は 5 であるから $NN(S')$ は減点される。

表 2: 各換言候補の $NN(S')$ $S$:意見を聞かない

| $S'$ | $p(S')$ | $r(S')$ | $NN(S')$ |
|---|---|---|---|
| 考えを押しつける | 2 | 3 | 0.7 |
| 独断で決定する | 2 | 3 | 0.7 |
| 病室へと歩き始める | 1 | 1 | 0.23 |

また、換言元動詞句と格助詞及び名詞の同一である係り元動詞句の動詞句対を対象として $NN(S')$ と同様に動詞-名詞頻度スコア $VN(S')$ を算出する。

#### 2.4.2 周辺単語一致度スコア

一般の文章において動詞句の周辺に出現する単語を周辺単語とする。周辺単語の集合を動詞句の周辺単語群とする。周辺単語の類似している動詞句同士は同様の文脈で利用されていることを意味している。このことから動詞句間の周辺単語の類似度は動詞句同士が換言表現であるかの指標として用いることができると考える。ここで動詞句間の周辺単語の類似とは以下の 2 点であると考える。

- 両動詞句の周辺単語の一致している数の多さ
- 両動詞句で一致している周辺単語の両動詞句周辺での出現頻度の高さ

これらを考慮して、換言元動詞句 $S$ とそれに対する 1 つの換言候補 $S'$ の間の周辺単語の一致度をスコアとして表す。これを周辺単語一致度スコア $WI$ として各 $S'$ に付与する。

対象となるコーパスから $S$ の周辺単語群 $W_S$ を抽出する。本稿では動詞句の周辺単語の抽出に Google を用いた。動詞句をクエリとして Google で完全一致検索したとき、検索結果ページに表示されるスニペット中に含まれるクエリ以外の単語を動詞句の周辺単語とした。検索エンジンを用いた理由については 3.1 節で述べる。また、周辺単語の品詞を「名詞-一般」、「名詞-サ変接続」とした。

$S$ 周辺における周辺単語 $w$ の出現頻度 $f_S(w)$ を求める。$f_S(w)$ の降順で $w$ に $W_S$ 内での順位 $r_S(w)(= 1, 2, 3, ...)$ を付与する。$S$ 周辺において複数単語の出現頻度が等しい場合の順位は、$S'$ 周辺での出現頻度の高い単語を上位とする。両動詞句周辺において出現頻度の等しい単語は同位とする。$W_S$ かつ $W_{S'}$ に含まれる $w$ について $r_S(w), r_{S'}(w)$ の逆数の相乗平均を算出する。この総和を $S$ と $S'$ の周辺単語一致度スコア $WI(S')$ とする (式 (3))。

$r_S(w), r_{S'}(w)$ が下位の $w$ が $WI(S')$ に与える影響は小さい。そこで本稿では $r_S(w), r_{S'}(w) \leq 30$ の $w$ を用いて $WI(S')$ を算出した。

$$WI(S') = \sum_{w \in (R_S \cap R_{S'})} 1/\sqrt{r_S(w) \cdot r_{S'}(w)}, (0 \leq WI) \quad (3)$$

表 3 で「病院」は換言元動詞句 $S$ 及び換言候補 $S'$ の周辺単語である。その出現頻度による順位は $S$ において 2 位,$S'$ において 3 位である。順位 $r_S(w), r_{S'}(w)$ の逆数の相乗平均を算出すると 0.41 である。「身体」「体調」についても同様とし算出される周辺単語一致度スコア $WI$ は 1.66 である。

表 3: 周辺単語の動詞句対における一致度

| $S$:体調を崩す | | | $S'$:風邪をひく | | |
|---|---|---|---|---|---|
| $w$ | $r_S(w)$ | $1/r_S(w)$ | $w$ | $f_{S'}(w)$ | $1/r_{S'}(w)$ |
| **身体** | 25 | 1 | **身体** | 32 | 1 |
| **病院** | 19 | 0.5 | のど | 25 | 0.5 |
| 精神 | 10 | 0.33 | **病院** | 14 | 0.33 |
| **体調** | 3 | 0.25 | **体調** | 9 | 0.25 |
| 調子 | 3 | 0.2 | 熱 | 5 | 0.2 |
| 会社 | 3 | 0.2 | 風邪 | 2 | 0.17 |

# 3 実験

提案手法を用いて換言元動詞句に対する換言動詞句の抽出実験を行った。実験には我々[6] が作成した Web コーパス約 9700 万文 (約 6.3GB) を用いた。そのうち約 8800 万文を用いて動詞句対を 650 万対収集した。これを本実験における動詞句対データとした。残りの Web コーパスを換言元動詞句の抽出に用いた。そこから約 650 万文抽出した。

## 3.1 周辺単語一致度スコアに関する実験条件

Web コーパスを用いて動詞句対データを収集していることで新聞コーパスには含まれていない表現を多く抽出していると考える。2 動詞句間の周辺単語一致度スコアを算出するにはコーパス中にその動詞句が含まれていなければならない。使用した Web コーパスは作成時に文の前後関係を考慮していないため本用途では用いることが出来ない。Web コーパスで抽出した表現の多くを包含するものとして Web がある。そこで Google による検索を行うことで擬似的に Web をコーパスとして用いた。本稿における動詞句の周辺とは Google のスニペットの範囲とした。動詞句の周辺の範囲は明確ではないため、検索からの処理が容易なスニペットを用いた。

一般的な単語は周辺単語として扱わない。Web コーパス中の「名詞-一般」と「名詞-サ変接続」で頻度の高いものから 500 単語をストップワードとした。

## 3.2 閾値

換言候補に付与した 3 つのスコアの閾値を決定するための実験を行った。収集した換言元動詞句より無作為に 200 文を 2 組抽出した。それぞれ閾値決定のための閾値用データセット、評価用データセットとする。

閾値用データセットの換言元動詞句に対する全ての換言候補を出力する。各候補が換言になり得るかを判定した。3 人の被験者が独立に判定を行った。独立に判定をしているので正解を多数決で判定する。

判定基準は「換言元動詞句を含む文において、動詞句を換言動詞句で置き換えたとき概ねの意味が保持される文があれば正解」とした。

換言候補が 1 つ以上の換言元動詞句は 200 文中 122 文であった。それに対する言い換え候補は全部で 1270 個であった。そのうち 220 個が正解、1050 個が不正解であった。この結果を利用して $NN$,$VN$,$WI$ の閾値を精度が最も高くなるようにする。

正解と判定する条件を次の 4 条件とし、各条件において最適な閾値を設定する。

(1)3 つのスコアが閾値以上, (2)2 つ以上のスコアが閾値以上かつ $WI$ が閾値以上, (3)2 つ以上のスコアが閾値以上, (4)1 つ以上のスコアが閾値以上

評価実験は閾値を定めた 4 条件で行う。そのとき最も適した条件での精度を本稿における実験結果とする。

表 4 に条件別の精度が最も良くなる各スコアの閾値およびそのときの適合率を示す。

表 4: 正解条件別の各スコアの閾値

| | $NN$ | $VN$ | $WI$ | 精度 | 適合率 |
|---|---|---|---|---|---|
| (1) | 0.91 | 0.84 | 0.53 | 85% | 69% |
| (2) | 0.81 | 0.92 | 1.17 | 84% | 65% |
| (3) | 0.99 | 0.95 | 1.14 | 83% | 52% |
| (4) | 0.99 | 0.99 | 1.97 | 77% | 32% |

# 4 結果

評価用データセットを用いて実験を行った。換言元動詞句に対して出力された換言動詞句を人手で評価した。3.2 節とは異なる被験者 3 人で評価方法については同様とした。正解条件を (2) としたときに出力された換言表現対は 239 組で適合率は 39%で最も高かった。そこで正解条件 (2) での結果を本稿における実験結果とする。抽出した換言表現を大きく 3 つのタイプに分けそれぞれ例に示す。換言表現対は「a」-「b」の形で表し、a は換言元動詞句、b は換言動詞句である。また換言表現対で換言可能な文と換言不可能な文例を示している。

**A．動詞句単位で換言可能**

換言しても原文の意味を保持出来る可能性のある動詞句対。

例 6)「知恵を出し合う」-「一緒に考える」

換言可： 良い解決策がないか 知恵を出し合う

換言不可：クラス全員の 知恵を出し合う

**B．内容語単位で換言可能**

換言しても原文の意味を保持出来る可能性のある内容語。

例 7)「感じを受ける」-「印象を受ける」

換言可：高貴な 感じを受ける

換言不可：痺れというよりも痛みに近い 感じを受ける

**C．誤り**

以下のような誤りが観察された。

・反意

例 8)「安打を放つ」-「安打を止める」

・関連

例 9)「社長を務める」-「会長に就任する」

・その他

例 10)「環境を整える」-「人が生活できる」

抽出した換言表現の各タイプの割合はA：30%, B：9%, C：61%である。本稿での目的であるAに分類される換言対の抽出は全体の 30%、正解出力の 78% であった。原文の概ねの意味を保持した換言の可能性をもつ動詞句対が抽出できている。

# 5 考察

## 5.1 換言表現の関係

本稿で収集した動詞句対は同じ係り受け関係であるが動詞句間の関係は一様ではない。換言表現となる動詞句間の

関係に特徴があるならば、その特徴に基づいた抽出が可能となる。そこで 2.2 節で動詞句対と共に収集した動詞句間を繋ぐ助詞をキーワードにその関係を考える。

表 5 に正解と判定された動詞句対 926 文での動詞句間を繋ぐ助詞とその割合上位 4 つを示す。

表 5: 動詞句対データ別の動詞句間を繋ぐ助詞の割合

| 対象 | none | て | とともに | たり |
|---|---|---|---|---|
| 正解動詞句対 | 52% | 29% | 3% | 3% |
| 不正解動詞句対 | 46% | 26% | 1% | 3% |
| 全動詞句対 | 43% | 28% | 1% | 3% |

各助詞は動詞句対を次の関係で繋いでいる。

- 「none(中止形)」並列
- 「て」因果、付帯状況・様態、並列
- 「とともに」並列、相関
- 「たり」並列

ただし「て」は「中止形」に置き換えられることから、「中止形」は「て」の関係を含んでいる。つまり正解単文においては上位 8 割、その他においては 7 割が並列、因果、付帯状況・様態の関係であると考えられる。

また表 5 には係り元動詞句の格助詞が「を」である全動詞句対及び不正解の 7909 文における各助詞の割合を示している。正解の動詞句対とそれ以外における各助詞の割合に大きな違いはみられない。換言表現となる場合に現れやすい助詞はない。よって換言表現となる動詞句間の関係に特徴があるとは言えない。

### 5.2 換言候補の数

換言元動詞句によって換言候補の数は大きく異なっている。換言候補の数が抽出精度に与える影響を考える。

評価データセットの換言元動詞句のうち換言候補の数が 50 個以上あるものに限って実験を行った。またその数を 100、150、200 以上と変化させた。換言候補の数と適合率の関係を表 6 に示す。これより換言候補数が多くなるにつれて出力結果の適合率が向上していることが確認できる。

表 6: 換言候補の数と適合率

| | $\geq$1 | $\geq$50 | $\geq$100 | $\geq$150 | $\geq$200 |
|---|---|---|---|---|---|
| 適合率 | 39% | 43% | 45% | 52% | 55% |

この結果が得られた理由としてまず、換言候補数が少ない場合は候補内に正解を含まないことがある。$NN$ 及び $VN$ ではそのスコアの付与の仕方から、候補数の少ないものに対しては減点を行っている。しかし、その減点規則が十分でないことが適合率の低さに影響を与えていると考える。

換言候補数が多くなると候補内に含んでいる正解数は増加する。しかし、それ以上に不正解数の増加がある。その中で候補数の増加に伴い適合率が向上していることから正解の判定を行うフィルタリングが機能していると考えられる。

各換言元動詞句のコーパス中での出現回数とそれに対する換言候補の数の相関を調べた。これより両者には 0.76 の正の相関があることがわかった。このことから動詞句対データを作成するコーパス量を増加させることでシステムの精度向上が期待できる。

### 5.3 不正解の文について

人手評価で換言でないと判定された動詞句を観察する。

不正解の動詞句は、正解の動詞句と比較して換言元動詞句と格助詞が一致していないものが多い。格助詞の変化により名詞と動詞の関係が大きく変化してしまう。そのため換言表現になりにくいと考える。

また、例 8, 例 9 のような反意、関連表現が誤りとしてみられる。これらは換言元動詞句と動詞句周辺単語が類似している可能性が高い。そのため $WI$ スコアが大きくなり正解と出力されたと考えられる。この場合シソーラスを用いるころで換言候補から除くことができると考える。

### 5.4 今後の課題

今後の課題として収集した換言表現の適用出来る箇所の特定がある。また、本研究の目的である読みやすい平易な文への換言表現を実現するためには文の読みやすさ、平易さの判定が必要である。

本手法は係り受け関係にある動詞句から換言表現対を抽出しているため 1 文中に共起する表現しか得られない。他の手法による抽出を目指すためにも換言可能な動詞句間の関係の検討は必要である。

## 6 まとめ

本稿では係り受け関係にある動詞句同士は換言表現となる可能性があるのではないかという観点から構文情報を用い、換言を獲得する手法を提案した。Web コーパスを用いた実験の結果 39%の精度で換言表現を抽出することが出来た。本稿で収集を目指した換言表現となる動詞句対は正解出力のうち 78%であった。また、このような換言表現は主にが並列、因果、付帯状況・様態関係から抽出できていることがわかった。


### 謝辞

本研究の一部は、科学研究費補助金 基盤 (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 によって実施した。


### 使用した言語資源及びツール

(1) 構文解析器 ‘‘CaboCha’’,Ver.0.52, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/˜taku/software/cabocha/

(2) 形態素解析器 ‘‘ChaSen’’,Ver.2.3.3, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/hiki/ChaSen/

### 参考文献


[1] 乾 健太郎, 藤田 篤: 換言技術に関する研究動向: 言語処理学会論文誌「自然言語処理」, Vol.11,No.5,pp.151–198,2004.

[2] 佐藤 理史: なぜ言い換え/パラフレーズを研究するのか: 言語処理学会第 7 回年次大会併設ワークショップ, pp. 1-2,2001.

[3] 弘前大学人文学部社会言語学研究室: 新版・災害が起こったときに外国人を助けるためのマニュアル: http://human.cc.hirosaki-u.ac.jp/kokugo/newmanual/top.html(2006/10/4).

[4] 仁井 康夫, 酒井 浩之, 吉田 辰巳, 増山 繁: 動詞間の換言知識の自動獲得: 言語処理学会第 9 回年次大会, pp.222-225,2003.

[5] 長谷川 隆明, 関根 聡: 教師なし学習による関係抽出に基づくパラフレーズの獲得: 情報処理学会 研究報告,NL-159-27,pp.193-200,2004.

[6] 関口 洋一, 山本 和英: Web コーパスの提案: 情報処理学会 研究報告,NL-157-17,pp.123-130,2003.