単文文型パターンの言い換えの抽出

小林 和晃　村上 仁一　徳久 雅人　池原 悟
鳥取大学 工学部 知能情報工学科
{kkobayas,murakami,tokuhisa,ikehara}@ike.tottori-u.ac.jp

1　はじめに

最近, 分かりやすい文への書き換えや要約や機械翻訳への適用を目指した言い換え技術に関する研究が盛んである. 言い換えのための知識 (事例や規則) としては, 語彙資源やコーパスからの収集が代表的である [5]. このうち, コーパスからの収集の方法において, [1] は対訳コーパスから同一の英文に対する複数の日本文を用意し, 日本文どうしのアライメントを取る手法を提案している. この手法は, 得られる日本文の組の数が少なく, かつ組内の日本文の文数も少ないと考えられる.

本研究は対訳コーパスから作成された文型パターンから, 同一の英語パターンに対する複数の日本語パターンを用意し, 日本語パターンのアライメントを取る手法を提案する.

今回は, 単文文型パターン [2] に提案手法を用い, 英語パターンが同一で日本語パターンが異なる日本語パターンの組を収集したときの, 日本語パターンの組の量を調査する. また, 得られた日本語パターンの組に対し, 言い換えの知識が収集できた割合を調査する.

2　対訳パターンを用いた言い換え知識の収集方法

2.1　原理

文型パターンは多くの日本文をカバーできるように変数によって表現されている. よって提案手法で日本語パターンの組を収集すれば, 従来手法よりも多く日本語パターンの組の数が収集できると思われる. したがって, 言い換えの知識を [1] の手法よりも多く収集できると考えられる.

2.2　特徴

提案手法は, 原文を変数に汎化することによって, 変数に当てはまる単語の言い換えが, 収集できない欠点があると思われる. しかし, 単語の言い換えはすでに同義語辞書が存在するため, わざわざ収集する必要性は低いと考えられる.

3　調査方法

3.1　調査対象

本研究では, CREST 対訳例文 100 万件 [3] から単文を収集して作成された単文文型パターン [2] を用い, 本手法の有効性を調査する. 以下に, 本研究で用いる単文文型パターンの概要を示す.

3.1.1　単文の条件

[2] で用いられた単文は, 基本的に文中に動詞が一つだけある文であり, 疑問文・命令文・会話文は対象外としている. 以下に例を示す.

(例 1)　彼は毎日自転車に乗る。

(例 2)　この林檎はややすっぱい。

3.1.2　単文文型パターンの例

以下に, [2] で作成された単文文型パターンの例を示す.

日本語原文：　妹は私と同じくらい一所懸命勉強する。
英語原文：　My sister studies as hard as I.
日本語パターン：　$N1$ は $PRO2$ と同じくらい一所懸命勉強する。
英語パターン：　My $N1$ studies as hard as $PRO2$.

単文文型パターンは, 7 つの日英対訳辞書を用いて原文中の対応関係が決定できた単語を N(名詞を表す) や PRO(代名詞を表す) のような変数に自動的に置き換えている. また, 日本語と英語の対応が取れるように変数に番号が付けられている.

3.2　言い換えの知識の収集方法

本研究では, 3.1.2 節で述べた単文文型パターン 215,342 件から, 提案手法を用い日本語パターンの組を 4,077 組収集した.

3.3　言い換えの知識の有無の調査

本研究では, 3.2 節で収集した 4,077 組のうちランダムで選んだ 100 組の日本語パターンに対し言い換えの知識

が存在するかを調査する. 調査は○・△・×の 3 段階で評価を行う. 以下に具体的に説明する.

○： 日本語パターンの組内の全ての日本語パターンにおいて言い換えの知識が存在

(例 3-1) $PRO1$ はよく $N2$ を 休む。

(例 3-2) $PRO1$ はよく $N2$ を 欠席する。

例 3 では, 例 3-1 と例 3-2 の両方から言い換えの知識が収集できるため, 評価○となる. なお, この例では「休む」と「欠席する」という言い換えの事例が収集できる. また, 例 3-1 と例 3-2 そのものが言い換えの規則としても収集できる.

△： 日本語パターンの組内の一部の日本語パターンにおいて言い替えの知識が存在

(例 4-1) $N1$ が $N2$ を押し流した。

(例 4-2) $N1$ が $N2$ を吹きとばした。

(例 4-3) $N1$ でその $N2$ が流された。

例 4 では, 例 4-1 と例 4-3 からは言い換えの知識が収集できるが, 例 4-2 では収集できないため評価△となる. なお, この例では言い換えの規則が収集できる.

×： 日本語パターンの組内の全ての日本語パターンにおいて言い換え知識が存在しない

4 調査結果

調査対象に提案手法を用いて収集した日本語パターンの組 4,077 組中ランダムで選んだ 100 組の日本語パターンに対し, 言い換えの知識が存在するかを調査した. 調査結果を表 1 に示す.

表 1 言い換えの知識の有無

言い換えの知識の有無	組数
○	63
△	8
×	29

この結果, およそ 7 割の組の日本語パターンには言い換えの知識が存在することが分かった. 以下に各評価の例を示す。

評価○の例

(例 6-1) $N1$ と $N2$ は相入れない概念だ。

(例 6-2) $N1$ は $N2$ と両立しない。

この例では日本語パターン全体が言い換え表現となっており, 文レベルでの言い換えの規則が収集できた.

評価△の例

(例 7-1) $PRO1$ は $N2$ がよい。

(例 7-2) $PRO1$ は 幸せな$N2$ の 下 に生まれた。

(例 7-3) $PRO1$ は 幸運な$N2$ の もと に生まれた。

この例では, 例 7-2 と例 7-3 から「幸せな」と「幸運な」および「下」と「もと」という言い換えの事例が収集できた.

評価×の例

(例 8-1) $N1$ は不在です。

(例 8-2) $N1$ 関節がはずれた。

この例では, 例 8-1 と例 8-2 の意味が異なるため評価×とした.

5 英語原文が同一の場合の言い換えの収集

5.1 調査目的

今回調査対象とした単文文型パターンには, 英語原文が一致する場合でも例 9 のように対応する日本語原文の表現がわずかに異なるために変数化の度合が異なり, 英語パターンが一致しない単文パターンの組合せがある.

前章までで調査を行った提案手法では, 英語パターンが同一で日本語パターンが異なるという条件で日本語パターンの組を収集しているため, 例 9 のような組合せは収集することができない.

そこでこの章では, 単文文型パターンから英語原文が同一で日本語パターンが異なるという条件で日本語パターンの組を収集する. また, 得られた日本語パターンの組内の日本語パターンに対し 3 章と同様に評価を行う.

(例 9-1)

日本語原文：道路がカチカチに凍っている。

英語原文：The road is frozen hard.

日本語パターン：$N1$ がカチカチに凍っている。

英語パターン：The $N1$ is frozen hard.

(例 9-2)

日本語パターン：道がかちかちに凍っている。

英語原文：The road is frozen hard.

日本語パターン：$N1$ が $ADV2$ に凍っている。

英語パターン：The $N1$ is frozen $ADV2$.

5.2 調査方法

単文文型パターン 215,342 件から, 5.1 節の手法を用い日本語パターンの組を 2,638 組収集した.

5.3 調査結果

英語原文が同一で, 日本語パターンが異なる日本語パターンの組 2,638 組中ランダムで選んだ 100 組に対し, 言い換えの知識が存在する日本語パターンの数を調査した. 調査結果を表 2 に示す.

表 2 英語原文が同一のときの言い換えの可能性

言い換えの可能性	組数
○	74
△	5
×	21

この結果, 5.1 節の手法で得た日本語パターンの組のうち, およそ 8 割の組の日本語パターンには言い換えの知識が存在することが分かった.

以下に各評価の例を示す。

評価○の例

(例 10-1) $PRO1$ は 難聴だ。
(例 10-2) $PRO1$ は 耳が不自由だ。

この例では,「難聴だ」と「耳が不自由だ」という言い換えの事例が収集できた

評価△の例

(例 11-1) $N1$ は きゅう覚 が $ADJ2$。
(例 11-2) $N1$ は鼻がよく利く。
(例 11-3) $N1$ は 嗅覚 が $ADJ2$。

この例では, 例 11-1 と例 11-3 からは言い換えの知識が収集できるが, 例 11-2 では収集できないため評価△とした. なお, この例では言い換えの規則が収集できた.

評価×の例

(例 12-1) その 会社は $N1$ を $VERB2$ た。
(例 12-2) 会社は $N1$ を $VERB2$ た。

この例では, 例 12-1 の日本語パターンしか比較要素が存在しないため評価×とした.

6 考察

6.1 英語原文の同一性の調査

今回対象とした日本語パターンの組には, 英語原文が同一の組も含まれている. そこで, 言い換えの知識が存在すると評価した日本語パターンの組のうち, 英語原文が同一である組の割合を調査した. 結果を表 3 に示す.

この結果, 提案手法は [1] の手法の倍程度の量の日本語パターンの組が収集できていることが分かる. しかし [1] の手法よりも言い換えの知識が存在する割合が低い. よって提案手法にさらなる制約を加えることが必要であると考えられる.

表 3 英語原文の同一性

評価	英語原文：同	英語原文：非同
○	41	22
△	4	4
×	5	24

6.2 言い換えの知識の種類の調査

言い換えの知識には, 言い換えの事例と言い換えの規則が存在する. そこで, 本研究で得た言い換えの知識が存在する日本語パターン 71 組に対し言い換えの知識の種類を調査した. 調査結果を表 4 に示す.

表 4 言い換えの知識の種類

言い換えの知識の種類	組数
規則	37
事例	34
計	71

この結果, 提案手法で得た言い換えの知識のおよそ半分は, 規則としての言い換えの知識であった. なお, 事例の言い換えの知識の内, 文そのものが言い換えの知識となっている組が 10 組存在した. 以下にそれぞれの言い換えの種類の例を示す.

言い換えの規則の例

(例 13-1) $N1$ が $N2$ で $VERB3$ た。
(例 13-2) $N1$ は $N2$ を $VERB3$ た。

言い換えの事例の例

(例 14-1) $PRO1$ は $PNOUN2$ $N3$ に たった。
(例 14-2) $PRO1$ は $PNOUN2$ $N3$ に 出発した。

文そのものが言い換えの知識の例

(例 15-1) われながら腑甲斐ないと思う。
(例 15-2) 赤面の至りです。

6.3 文型パターン作成方法との関係

提案手法は, 文型パターンの作成方法によって収集結果が変わると考えられる. 例 8 で言い換えの知識が存在

せず×と評価した例の単文パターンの詳細を例 16 に示す. 例 16-2 を見ると, 本来「肩関節」までが変数化され $N1$ となるべきはずが「肩」しか変数化されていない.

この理由として, 今回用いた単文文型パターンは, 変数化を自動的に行ったため変数化が正しく行われていないためだと考えられる. 人手により正しく変数化が行われれば, より言い換えの知識を収集できると考えられる.

(例 16-1)

日本語パターン：$N1$ は不在です。
日本語原文：タバコが無くなった。
英語原文：My tobacco is out.
英語パターン：My $N1$ is out.

(例 16-2)

日本語パターン：$N1$ 関節がはずれた。
日本語原文：肩関節がはずれた。
英語原文：My shoulder is out.
英語パターン：My $N1$ is out.

また, 文型パターンを作成する際に名詞や用言に対し, 名詞や用言の意味属性 [4] を付与することによって, より信頼性の高い言い換えの知識が獲得できると考えられる.

7　おわりに

本研究では, 言い換えの知識を収集するため, 対訳コーパスから作成された文型パターンから, 同一の英語パターンに対する複数の日本語パターンを用意し, 日本語パターンのアライメントを取る手法を提案した.

提案手法の効果をはかるため, 単文文型パターン 215,342 件から提案手法によって日本語パターンの組を 4,077 組収集した. また, 収集した日本語パターンの組に言い換えの知識が含まれている割合を調査した. 4,077 組からランダムで 100 組を選び, その日本語パターンに言い換えの知識が含まれているかを調べた結果, 71 組の日本語パターンは言い換えの知識が含まれていることが分かった.

今後の予定として, 本研究で収集した全ての日本語パターンに対し調査を行っていきたいと考えている.

謝辞

本研究は，科学技術振興事業団「JST」の戦略的基礎研究推進事業「CREST」における研究領域「高度メディア社会の生活情報技術」の研究課題「セマンティックタイポロジーによる言語の等価変換と生成技術」の支援によるものである．研究の助言を頂いた研究室メンバーに感謝する.

参考文献

[1] 乾:言語表現を言い換える技術, 言語処理学会第 8 回年次大会チュートリアル, pp.1-21, 2002.
[2] 西山ほか:単文文型パターン辞書の構築, 言語処理学会第 11 回年次大会発表論文集, pp.372-375, 2005.
[3] 村上ほか:日本語英語の文対応の対訳データベース, 「言語・認識・表現」研究会, 2002.
[4] 池原ほか:「日本語語彙大系」, ISBN4-00-130101-6, 岩波書店, 1997.
[5] 今村ほか:階層的句アライメントを用いた日本語翻訳文の換言, 言語処理学会第 7 回年次大会ワークショップ, pp.15-20, 2001.