形容詞述語文の換言事例の分析

大竹 清敬　　山本和英
ATR音声言語コミュニケーション研究所
{kiyonori.ohtake, kazuhide.yamamoto}@atr.co.jp

概要

旅行会話を集めた2言語パラレルコーパスから得られた換言事例を分析する．末尾の文節に形容詞を含む文から末尾文節に形容詞を含まない文への換言事例を対象とする．

1 はじめに

換言処理は，自然言語処理において非常に重要な処理である．一般に，あることを表現するために，様々な言い方が存在する．そして，これは自然言語が持つ重要な特徴の一つである．しかしながら，同時にそれは自然言語処理において一つの問題でもある．

我々は，自然言語処理の一つの応用として音声翻訳を考え，研究を進めている．既存の機械翻訳の枠組で翻訳の精度を向上させるためには，規則を網羅的に整備する，さまざまな状況で用いられる用例を収集するなどの作業が必要となる．一方で，そのような作業を適切にすることができる人間は限られている．そこで，我々は，単言語内での処理に重点を置き，原言語と目的言語の双方で換言処理を行うことで変換処理の負荷を最小限に抑えた翻訳方式 [1] について研究を進めている．

単言語における換言処理に重点を移すことによって，翻訳に利用できる資源が増える．たとえば，日本語の文から適切な英語訳を作成できる人間よりも，同等の意味を持つ文に適切に換言できる人間のほうがはるかに多い．そしてこの人数は，日本語からどの言語へ翻訳する場合にも，変わらずほぼ一定である．

換言処理は同一言語における翻訳と考えることができる．そのため，従来の機械翻訳同様にいくつかのアプローチが可能である．換言現象が比較的局所的な場合，人手で規則を記述し，処理することが実用的である事を示した [2, 3]．同時に，文献 [3] では，規則による換言処理のみでは対処しにくい現象があることについて触れている．そこでは，事例に基づく換言がそのような現象に適していると述べている．

事例に基づく実用的な換言を実現するためには，大量の換言事例が必要となる．換言事例の収集は，翻訳事例の収集よりコストが少ないと予想できる．しかし，問題となるのは，何のために換言するのかという目的（換言因子 [4]）である．無目的に換言事例の収集を行う場合，ある程度の換言事例を容易に収集できると予想する．しかし，収集した換言事例を有効に利用できるかどうかの点は疑問である．逆に，ある換言因子が極端に厳しい条件を要求する場合，換言事例の収集が困難になると予想する．今回の報告では，機械翻訳への適用を目的として，形容詞述語文から非形容詞述語文への換言について分析する．

大量の多言語パラレルコーパスが存在したとき，そこから換言事例を自動的に抽出できる [5]．本研究では，旅行会話を集めた日英のパラレルコーパスが与えられたときに，そこから得られる換言事例について分析する．本研究の目的は，(a) パラレルコーパスから自動的に得られた事例がどの程度，換言処理に利用できるのか，(b) それらの換言事例にどのような換言処理を実現できるのかを明らかにすることである．

2 換言事例の収集

旅行会話を集めた日英のパラレルコーパス (以下，パラレルコーパスとする) を用意し，そこから換言事例を抽出する．パラレルコーパスはのべ162320文からなる．1文毎に対訳が対応づけられている．英語文の異なりは98290であり，日本語文の異なりは102664である．

基本的な換言事例の収集方法は，新しいものではなく，Barzilay らの方法 [5] と同様である．それは，ある文に対して複数の対訳が存在するとき，それらの対訳を換言事例とみなす手法である．我々が使用するパラレルコーパスは文単位での対応がとれており Barzilay らが対象とした小説と異なり，換言事例を収集するために考慮しなければならないこと (対応づけの必要性やスタイルの違いに対する処理) は少ない．

具体的な収集方法は，ある英語の文 E_i に対して，複数の日本語訳 $J_{i1}, \cdots J_{im}$ が存在するとき，可能な2文の組み合わせ J_{ij} と J_{ik} を換言事例とする．このとき，$1 \leq j, k \leq m, j \neq k$ である．そしてこれを全ての英語

文について網羅的に収集する．一方，一度収集された換言事例に含まれる日本語文（たとえば，J_{ij}）に対応する英語の文 $(E_i, \cdots, E_n)$ を求め，それらの日本語訳を再度集めることによってより多くの換言事例を収集できる可能性がある．しかし，この操作によって換言事例としてはふさわしくない事例も多く集める可能性がある．そのため，今回の報告では，ある英語の文に対する日本語訳が複数存在する場合に限定して換言事例を収集した．つまり，再帰的に換言事例を収集することは行っていない．この方法によってパラレルコーパスから収集された換言事例は，全部で 102747 組である．

3 形容詞述語文

本研究における形容詞述語文の具体的な定義は，普遍的な形容詞述語文の定義が困難であること，ならびに再現性を考慮して，実際に使用した解析器の品詞定義によるものとする．我々は，形態素解析器として ChaSen[1]，構文解析器として CaboCha[2] を使用した．それらの解析器を用いた本研究における形容詞述語文の定義は「末尾の文節に形容詞-* [3]または名詞-形容動詞語幹 の品詞を持つ形態素を含む文」である．

形容詞述語文には，たとえば，「いえ結構です」のようないわゆるダ文が含まれる．また，「…いいですか」などの述語も形容詞述語文に含まれる．これらの述語は，さまざまな状況で使用することができる．そのため，一つの述語でさまざまな意味を表す．したがって，機械翻訳をはじめとする自然言語処理全般において形容詞述語文の換言の分析は重要である．

形容詞述語文の一部 (…大丈夫ですか，…いいですか，など) は一つの表現でさまざまな意味を伝える．そのため，これらの表現は非常に利用しやすく，会話でも多用される．たとえば，「…いいですか」の機能的役割を見てみると，

- 確認
  例) これでいいですか．
- 許可を求める
  例) タバコを吸ってもいいですか．
- 方法をたずねる
  例) どのように書けばいいですか．
- 好みをたずねる
  例) どんな色がいいですか．

などのように分類できる．

[1] http://chasen.aist-nara.ac.jp/
[2] http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/
[3] つまり形容詞-で始まる品詞全てを指している．

規則に基づく翻訳手法によってこれらの文を翻訳することを考えた場合，網羅的に規則を収集しなければ実用的な翻訳を達成できない．また，用例に基づく機械翻訳手法であっても，対応する用例を持たなければ適切に翻訳することは難しい．たとえば，市販されている O 社の日英機械翻訳ソフトで上の例の一部を翻訳してみると「タバコを吸ってもいいですか．/May I smoke?」，「どのように書けばいいですか．/How are you good if you write it?」のようになる．我々はこの機械翻訳ソフトの翻訳手法を知らないが，どのような手法であっても，一つの表現でさまざまな意味を表す述語に対する処理が難しいことを示している．

4 文脈依存性の分析

収集された換言事例 102747 組から，本研究における形容詞述語文から非形容詞述語文への換言事例 13323 組を抽出した．また，同一の形容詞述語文から複数の非形容詞述語文へ換言される場合がある．これらの換言事例の数が多すぎる場合，いくつかの問題が考えられたので制限した．

一つめの問題は，実際に換言処理を適用する際に換言の候補が多くなりすぎて処理が繁雑になることである．もう一つの問題は，大量の換言候補を持つ場合の多くは機能語の違いによって生じており，換言事例としての重要度が低いことである．具体的な制限として，一つの形容詞述語文から非形容詞述語文への換言事例が平均で 2 以下となるようにある閾値以上の換言事例を除いた．結果として一つの形容詞述語文から非形容詞述語文への換言事例が 5 以下のものを対象としたとき，平均が 2 以下となった．このような換言事例は 3660 個あり，その換言対象となった形容詞述語文は 1879 文である．また，除かれた換言事例 9663 を構成する形容詞述語文は 638 文である．

4.1 文脈依存性に基づく分類

得られた 3660 個の換言事例が換言処理にどの程度用いることができるのかを調べるために換言事例を次の 3 つのクラスに分類した．

(A) 文脈に依存せず，いつでも換言事例を適用できる．

(B) 換言事例の適用が文脈に依存する．つまり言い替えると，適用すると意味をとりちがえてしまう文脈が存在する．

(C) 不適切な換言事例．つまり換言事例ではない．

換言事例 3660 個をこれらの 3 つのクラス分類した結果を表 1 に示す．

表 1: 換言事例の文脈依存性の分類

クラス	(A)	(B)	(C)
事例数	2068	1258	334
全体に対する割合	56.50%	34.37%	9.13%

4.2 文脈依存性に関する考察

表 1 から分かるのは，半数以上の換言事例を事例に基づく換言処理にそのまま適用できるということである．しかしながら，クラス (A) に属する事例を自動的に識別することは容易ではない．そのため，パラレルコーパスから収集した事例に基づく換言処理を考える場合には，クラス (B)(C) に属する事例をいかに排除するかが問題となる．

クラス (B)(C) に属する事例を排除するために，換言事例を収集する段階で，制限する方法がある．今回報告した収集方法ではある英語の表現 E_i の対訳が複数存在したとき $J_{i1}, \cdots J_{im}$，これらの全ての組み合わせを換言事例としている．そうではなく，m の大きさや個々の日本語表現のパラレルコーパス中での頻度などを考慮し，ある尤度を導入することにより不適切な表現を制限できると予想する．

また，今回収集した換言事例の中の 1 割近くがクラス (C) に属する結果となった．この理由の一つは，文脈の情報を無視して換言事例を収集したところにある．我々が用いたパラレルコーパスには，それぞれの文が使用される文脈が何らかの形で記述されている．たとえば,「空港で」や「ショッピングで」などのキーワードによって表示される．しかし，現段階では，これらのデータが機械可読の形式になっていない．そのため，文脈によって内容が変化する同一表記異内容の表現から換言事例として不適切なものが抽出された．このことから，これらの文脈情報を用いることができれば，クラス (B) や (C) に属する事例を自動判別できるばかりでなく，クラス (B) に属する事例を活かした換言処理が可能になる．

ここで，文脈情報を用いた自動判別の一例を与える．まず，ある文 J_{ij} の文脈 $C(J_{ij})$ が単語で与えられるとする．換言事例を $J_{i\alpha} \rightarrow J_{i\beta}$ と表記する．さらに $C(J_{i\alpha}) = C(J_{i\beta})$ であるとする．このとき，$J_{i\alpha}$ あるいは $J_{i\beta}$ に表記が非常に類似した $J_{l\gamma}$ があり $C(J_{i\alpha}) \neq C(J_{l\gamma})$ あるいは $C(J_{i\beta}) \neq C(J_{l\gamma})$ ならば 換言事例 $J_{i\alpha} \rightarrow J_{i\beta}$ がクラス (B) に属する可能性が高まる．また，$C(J_{i\alpha}) \neq C(J_{i\beta})$ ならば，それはクラス (C) に属する事例である可能性が高まる．

クラス (B) に属する換言事例を活用するためには，適用しようとする文 K の文脈 $C(K)$ を単語で与える必要がある．もしこれが適切に与えられるならば，K に適用しようとする換言事例 $K \rightarrow J_{i\alpha}$ が適用可能かどうかは，$C(K)$ と $C(J_{i\alpha})$ が同一かどうかでわかる．

5 換言事例の分析

クラス (A) に分類された 2068 の換言事例は，文脈に依存せず換言処理に用いることができる．これらの事例によってどのような換言処理が実現できるか分析した．

この報告では，換言処理を大きく 3 つに分けて考える．(1) 語彙的換言処理 (2) 構文的換言処理 (3) 意味的換言処理 である．(1) 語彙的換言処理とは，表記の違いや (財布⇔サイフ)，動詞の機能的表現の違い (X してください⇔X してくださいませんか) などの局所的変化を扱う換言である．(2) 構文的換言処理は，語順の入れ換え (どのくらいの高さ⇔高さはどのくらい) や，助詞の交替を伴う述部の換言 (カットだけでいいです⇔カットだけをお願いします) などを扱う．(3) 意味的換言処理は，(1) および (2) の換言処理を併用しても実現できない換言を扱う．

我々は既に，換言現象が比較的局所的な場合，人手で規則を記述し，処理することが実用的である事を示している [2, 3]．したがって，パラレルコーパスから得られたこれら 2068 の換言事例には，局所的な換言現象ではなく，より大きな単位での換言が含まれていることを期待している．

5.1 換言事例の類別

文脈に依存せず用いることができる換言事例について，どのような換言処理を適用するとその事例を再現できるかという観点からこれらの事例に 3 つの素性を付与した．3 つの素性を以下に示す．

(L) 語彙的な換言．換言されている部分の前後の形態素程度の条件でその換言を実現できる．

例) お会いできて嬉しい．
→ お会いできてうれしく思います．
パスポートをなくしてしまいました．
→ パスポートを紛失してしまいました．

(S) 構文的な換言．語順の入れ替えや，態の変換，さらに用言相当句の換言による助詞の交替をともなう換言なども含める．

例) どのくらい深いですか．
→ 深さはどのくらいですか．

コーヒーが欲しいのですが.
→ コーヒーをください.

(E) (L) と (S) を組合せても実現できない換言.

例) いえ結構です. → お断りします.
そこまでは遠いのでしょうか.
→ ここから離れていますか.
それ以上は安くなりませんか.
→ それが最終的な値段ですか.
どこで払えばいいですか.
→ 支払い場所はどこですか.

これらの素性に基づき，文脈に依存しない2068の換言事例を (L) のみ，(S) のみ，(L) と (S) 両方，(E) のみをそれぞれ持つ4つのクラスへ分類した結果を表2に示す.

表 2: 文脈に依存しない換言事例の分類

素性	(L)	(S)	(L)+(S)	(E)
事例数	629	355	330	754
割合	30.42%	17.17%	15.96%	36.46%

5.2 換言事例の類別に関する考察

素性 (L) を持つ換言事例は，その変化が局所的であることから，規則の記述が比較的容易であり，規則によって実現できる．また素性 (S) を持つ換言事例も，動詞に関する辞書 (たとえば格フレーム) などの整備が必要であるが，その換言は同様に規則によって実現が可能である．一方，素性 (E) を持つ換言事例から規則を抽出することは困難である．したがって，これらの事例を数多く収集できれば，換言処理に幅を持たせることができる．しかしながら実際に抽出した換言事例の約6割は (L) または (S) の素性を持つ．換言因子にも依存するが，我々が用いたパラレルコーパスから素性 (E) を持つ事例を大量に収集できるとは断言できない.

コーパスから素性 (E) を持つ事例を集めることが容易ではないとすると，人手で作成する方法と比較検討する必要がある．素性 (E) を持つ換言事例には，短時間で作成できそうにないものも含まれる．たとえば，“もう少し安くなりませんか” に対する “それが最終的な値段ですか” などである．したがって，これらの換言事例を人手で作成する作業は，容易ではないと予想する．一方，収集された事例には $J_{ij} \rightarrow J_{ik}$ が素性 (E) を持ち，$J_{ij} \rightarrow J_{il}$ が素性 (L) または (S) を持つ場合が存在する．たとえば，“すわってもいいですか. → この席はあいていますか. ” は素性 (E) を持つが，“すわってもいいですか. → すわってもかまいませんか. ” は素性 (L) を持つ．また，人間がこれらの素性 (L)，(S) を組みあわせた事例を作成することは比較的容易だと予想する．このことから，ある換言因子を満たすためだけならば素性 (L) または (S) を組み合わせた換言で十分対応できる可能性がある.

また，素性 (E) を持つ換言事例の一部は，副次的である．$S_a \rightarrow S_b$ という換言事例が素性 (E) を持っていたときに，S_b に対して (L) または (S) の素性を持つ換言を適用して得られた S'_b によって $S_a \rightarrow S'_b$ という素性 (E) を持つ換言事例が得られる．したがって，本質的に素性 (E) を持つ事例は表2に示される数より小さい.

6 むすび

日英パラレルコーパスから換言事例を収集し，形容詞述語文から非形容詞述語文への換言事例について分析した．単純に自動収集した事例の約4割は，文脈に依存するもの，事例として不適切なものであった．自動抽出した事例を換言処理に適用するためには，尤度による不適切な事例の制限などが必須である．また，文脈に依存せず適用できる換言事例の6割は規則化が可能であり，残り約4割は我々が望む規則化が困難な事例であるが，副次的な事例も含まれる.

本研究は通信・放送機構の研究委託により実施したものである.

参考文献

[1] Yamamoto, K., Shirai, S., Sakamoto, M. and Zhang, Y.: Sandglass: Twin paraphrasing spoken language translation, in *Proceedings of ICCPOL 2001*, pp. 154–159 (2001).

[2] Yamamoto, K.: Paraphrasing Spoken Japanese for Untangling Bilingual Transfer, in *Proceedings of NLPRS2001*, pp. 203–210 (2001).

[3] Ohtake, K. and Yamamoto, K.: Paraphrasing Honorifics, in *Workshop Proceedings of Automatic Paraphrasing: Theories and Applications (NLPRS2001 Post-Conference Workshop)*, pp. 13–20 (2001).

[4] 山本和英：換言処理の現状と課題, 言語処理学会第7回年次大会ワークショップ論文集, pp. 93–96 (2001).

[5] Barzilay, R. and McKeown, K. R.: Extracting Paraphrases from a Parallel Corpus, in *Proceedings of ACL 2001*, pp. 50–57 (2001).