換言コーパスを利用した中国語換言処理

張 玉潔　山本 和英　坂本 仁
E-mail: {yzhang, yamamoto, msakamo}@slt.atr.co.jp
ATR 音声言語コミュニケーション研究所

1. はじめに

我々は換言処理 (paraphrasing) に重点を置いた音声翻訳手法を提案した [Yam01]。日常会話では、しばしば相手が分からない言葉を別の言い方にする。これにならい、機械翻訳では、原言語の発話が翻訳できない場合、翻訳できそうな言い方に変えようという考え方である。本稿は中日音声翻訳システムにおける中国語換言処理 [Zha01a, Zha01b] に関して、その最新の研究結果を報告する。

2. 換言処理の目標と手法

翻訳の前処理として、中国語換言処理は次の目標を目指す。

(1) 口語表現をフォーマルな表現に整える。
(2) 構文、意味上の曖昧さを減少させる。
(3) 変換処理が翻訳できる表現をカバーするにより多く表現を生成する。
(4) 全文の換言により何も翻訳出来ない場合、情報を落として換言する。

上記の (3) は表現を多様化するための換言であり、(1)、(2) と (4) は表現を簡単化するための換言と考えてもよい。現在、(1)、(2) と (3) について研究している。

換言処理は、同じ言語の中で、入力文の意味をできるだけ保存するように別の文を自動的に生成するという処理である。一見すると、入力文を解析し意味表現を得て、その意味表現から文を生成するというプロセスになって、自然言語処理の意味解析と文の生成という二つの問題に帰着することになる。ところが、以下の理由により、このような考え方は妥当ではない。

(a) 現在、中国語の構文解析と意味解析の技術はまだ使えるレベルになっていない。特に、話し言葉を対象とする研究は始まったばかりである。構文解析と意味解析自体はそれぞれ大きな研究課題になっている。換言処理では、構文および意味解析がどこまで必要かを明らかにした上で、対応を取るべきと考える。

(b) たとえ意味表現を得ても、ただ一つの表現を生成すると、(3) の目標を達成できない。換言処理では多様な表現を生成できることは最も重要である。この点は従来の生成と異なる。

実際、換言処理は語彙レベル、句レベル、構文レベルなど様々なレベルで行うことができる。ある程度の文脈と関連するが、必ずしも全文の意味を取らないとできないわけではない [Kat00]。これらの理由で、我々は当面形態素解析のみを行い、パターンに基づく手法で換言処理を行うことにした。研究の重点は多様な表現を生成できることに置き、換言コーパスから換言パターンを獲得する手法を提案した。

3. 換言パターンの獲得

ATR の旅行会話の中国語換言コーパス [Zha01b] は、2 万文の原文と 4 万文の換言文からなる。一つの原文には二つの換言文が対応する。換言現象の種類は語順の入れ替え、同義語と他の文型表現の取り替えなどがある。4 万文の換言文のそれぞれと対応する原文との組を換言文対と呼び、全部で 4 万の換言文対がある。4 万の換言文対に対して、単語分割と品詞付与を行った。品詞体系は Penn Chinese Treebank の品詞分類を採用した [Xia00]。換言パターンの獲得はこの形態素解析済みの換言文対の上で行う。

3.1 関連の換言文対の抽出

換言パターンの獲得は換言現象を分けて行う。これにより、換言処理は文をある方向へ換言させることが制御でき、また、その方向への換言によりどのような情報が変わるかを整理することができると期待する。今まで次のような換言現象に関して換言しようと考えて、関連の換言文対を抽出した。

3.1.1 語順の入れ替え

中国語の口語では語順がかなり自由である。そこで語順を入れ替えて換言する研究を行った [Zha01b]。ここでは、まず、原文と換言文の単語数が同じであり、原文の任意の単語が換言文にあり、かつ換言文の任意の単語が原文にあるような換言文対を抽出した。例1は語順に関する換言文対の例である。抽出した換言文対は一方の例文が他方の例文の語順に従って入れ替えれば換言文が得られるという知識を含んでいる。

例1

原文: 有没有 我的留言

(私宛ての伝言は届いていませんか)

換言文: 我的留言 有没有

3.1.2 否定表現

換言文対を観察した結果、原文には“不”と“没”の否定表現を含んで、換言文には肯定表現になっているという現象がある。例えば、原文の反復疑問、二重否定などの表現は換言文において肯定表現に言い換えられる。換言処理でこのような否定表現を肯定表現に言い換えれば、翻訳の変換部にとってきっと有益であろう。そこで換言文対から、一方の文が“不”あるいは“没”を含んで、他方の文が含んでいないような文対を抽出した。例2は否定表現に関する換言文対の例である。

例2

原文: 知不知道 我的電話

(私の電話を知っているか知っていないですか)

換言文: 知道 我的電話吗 (私の電話を知っていますか)

3.1.3 文法標識“把”

中国語では表層的な文法標識が少ないが、前置詞“把”はその一つである [Zha99]。“S (主語) V (動詞) O (目的語) C (補語) ”のような普通語順に対して、“把”を入れて“S 把 O V C”の形に言い換えることが可能である。“把”は目的語を動詞の前に移動し、目的語を強調する役割がある。構文上では、“把”があると目的語を特定することがより容易にできる。よって、“把”を入れることは構文上の情報を増やすことになる。また、“把”を含むことにより目的語を特定出来て、他の表現に言い換えるときより精確になると予想する。換言文対から、一方の文が“把”を含んでおらず、他方の文が含んでいるような換言文対を抽出した。例3は“把”構文に関する換言文対の例である。

例3

原文: 这 /DT 张 /M 单子 /NN 请 /VV 您 /PN 填好 /VV (この用紙ご記入頂けますか)

換言文: 请 /VV 您 /PN 把 /BA 这 /DT 张 /M 单子 /NN 填好 /VV (この用紙にご記入頂けますか)

3.2 換言文対の抽象化

次に抽出した換言文対を抽象化して換言パターンを得る。換言文対の一方の例文から換言パターンの条件部分を、他方の例文から生成部分を得て、換言パターンを構成する。換言パターンを適用するとき、入力文が条件部分に適格であれば、生成部分にしたがって換言文を生成する。

換言処理は意味を保持することが重要である。例文をどこまで抽象化して適当かは例文による。基本的に、形態素の品詞は構文上の情報を持つから、すべて保持する。表記については、動詞、助動詞、副詞、前置詞、量詞、語気を表す助詞が文の大意を定めるので、それらの表記は保持して、名詞、代名詞、数詞の表記は変数で置き換えて抽象化する。例えば、例3の換言文対を抽象化して例4のパターンが得られる。

例4

这 /DT 张 /M X_1/NN 请 /VV X_2/PN 填好 /VV →请 /VV X_2/PN 把 /BA 这 /DT 张 /M X_1/NN 填好 /VV

X_i は任意の表記にマッチできることを意味する。しかし、このような自動的な抽象化には次の問題があった。まず、長い例文から得たパターンには、形態素の表記が多すぎ、それにマッチ出来る文が少ない。実際、文の換言とあまり関係のない部分、例えば、名詞を修飾するような節があり、それをさらに抽象化できると考える。次に、名詞と代名詞の中で、いくつかの形態素が特別な構文成分と意味を持つため、それらの表記を抽象化すべきでない。例えば、動詞の量詞として使われる名詞“一下”(ちょっと)とか、疑問の意味を持つ数詞“几个”(何個)などがある。このような形態素の表記は保持すべきである。

抽象化の可否の情報は人の知識からしか得ることができない。自動的な抽象化に人の知識を組

み込むために、換言文対を編集するツールを開発した。ツールは編集用符号と転換プログラムから構成される。編集用符号は人がそれを用いて換言文対の上で抽象化の範囲を定義する。転換プログラムは定義された換言文対から換言パターンを得る。編集用符号は三つあり、次のように定義される。

[ ]: 形態素列を囲む。転換では、最後の形態素の品詞を除いて、囲まれた部分を変数で置き換える。これにより、囲まれた部分は抽象化され、その部分の構文上の役割は最後の形態素の品詞により保持される。

{ }: 一つの形態素を囲む。形態素の表記には一つ以上の値を定義できる。転換では、囲まれた部分の表記は保持される。一つ以上の表記を定義できることにより、同じ意味の単語、あるいは、同じように換言出来る単語が一つのパターンに収まることが出来る。

< >: 一つの形態素を囲む。転換では、その表記が変数で置き換えられる。これにより、ある文脈で品詞が動詞、副詞であるような形態素も抽象化できるようになる。

例をあげると、例5のように換言文対が定義され、例6のパターンが得られる。

例5

原文: 请/VV 给/VV 我/PN 两/CD (本/M) [日语/NN 的/DEG] [指南/NN 手册/NN]
(二部の日本語の案内書をください)
換言文: [日语/NN 的/DEG] [指南/NN 手册/NN] 请/VV 给/VV 我/PN 两/CD (本/M)
(日本語の案内書を二部ください)

例6

请/VV 给/VV X_1/PN X_2/CD X_3/M Y_1/DEG Y_2/NN → Y_1/DEG Y_2/NN 请/VV 给/VV X_1/PN X_2/CD X_3/M

X_i は例4と同じであり、Y_i/POS は最後の形態素の品詞が POS であるような形態素列にマッチできることを意味する。

3.3 換言パターンの作成

編集ツールを用いて、抽出した換言文対から次の四種類の換言パターンを得た。

(1) 否定表現の削除パターン: 459個
(2) "把"の挿入パターン: 160個
(3) "把"の削除パターン: 160個
(4) 語順の入れ替えパターン: 2030個

このうち (3) は (2) のパターンの条件部分と生成部分を逆にしたものである。

4. 換言処理の制御

より多く表現を生成するために、多種の換言パターンをどういう順番で適用するかについて述べる。2節で述べたように、換言処理は、表現を簡単化するための換言と表現を多様化するための換言がある。簡単化の換言は構文と意味の曖昧さを減少させるから、そのあと多様化の換言をすると、より正しい換言文が得られると予想する。よって、簡単化の換言を先に行うべきである。3.3で得られた換言パターンのうち、(1) の換言パターンは簡単化の換言であり、他の (2)、(3) と (4) の三種類の換言パターンは多様化の換言である。(1)、(2)、(3)、(4) の順番にパターンの適用を行い、次のような手続きで換言する。

(i) 入力文をすべての種類のパターンの適用データにする。
(ii) 毎種類のすべてのパターンを適用データに適用してみる。適格なら換言文を生成する。生成した換言文を後ろのすべての種類のパターンの適用データにする。
(iii) 変換部に換言結果を渡すとき、同じ換言文を二度出さないようにする。

これにより、換言処理の任意時点で、そこまで生成した換言文を変換部に渡すことができる。変換部が換言文を翻訳できれば、換言処理が停止する。また、得られた換言文をさらに換言することができ、より多く表現を生成できることを期待する。

5. 換言実験と評価

得られた換言パターンを用いて換言コーパスの上で換言実験を行った。オープン実験の効果を得るために、入力文にはその文から得られた換言パターンを適用しないようにした。生成した換言文は人手により評価した。換言文は文として正しく、かつ意味的に入力文と同じなら、正解とする。評価結果を表1に示す。

表1により、入力文中の4908文、約10.9%が換言された。換言された文において、平均で一文に1.66個の換言文、最大で一文に四つの換言文

表 1. 換言実験の結果

入力文	45110
うち換言された文	4908 (10.9%)
生成した換言文	8183 (1.66 文)
うち正しい換言文	7226 (88%)

が生成された。生成した換言文は約 88% の正解率を得た。換言結果の二つの例を上げる。

例 7

入力文: 能不能 帮我订到最早的班机呢? (予約できる一番早い便をとってもらえるか)

換言文 1: 能 帮我订到最早的班机吗?

換言文 2: 可以 帮我订到最早的班机吗?

例 8

入力文: 风景漂亮的房间 请给我 。 (景色が綺麗な部屋下さい。)

換言文 1: 请 把 风景漂亮的房间给我。(景色が綺麗な部屋を下さい。)

換言文 2: 我 想要 风景漂亮的房间。 (景色の綺麗な部屋がほしい。)

換言文 3: 请给我 风景漂亮的房间。 (景色の綺麗な部屋を下さい。)

換言文 4: 房间 请给我 风景漂亮的。 (部屋は景色が綺麗なほうを下さい。)

例 7 の入力文には反復疑問表現 "能不能" を含んでいる。否定表現の削除パターンを適用した結果、換言文 1 と換言文 2 が正解で、肯定の表現になっている。例 8 の入力文には、目的語 "风景漂亮的房间" と述語 "请给我" の語順は倒置になっている。入力文に "把" の挿入パターンを適用することにより、換言文 1 が得られ、普通の語順になっている。換言文 1 に "把" の削除パターンを適用することにより、換言文 2 と換言文 3 が得られた。換言文 3 に語順の入れ替えパターンを適用することにより、換言文 4 が得られた。換言文 1、2、3、4 のいずれも正解である。これらの結果から、提案した換言手法は高い正解率で文の簡単化と多様化へ換言できたことが分かった。

生成した換言文の誤りを分析した結果、一つの原因は換言パターンによりある成分を取り出すには誤りがあったからである。例えば、"请给我两个人住的房间" (二人の部屋をください) にパターン例 6 を適用すると、数量句 "两个" は "人" を修飾しているのに、"人住的房间" を修飾していると間違えて、"人住的房间请给我两个。" (人が住んでいる部屋を二つください) のように換言してしまった。もう一つの原因は形態素解析の品詞付与誤りである。例えば、"请给我订机票" (チケットの予約をしてください) において、"给" が前置詞であるのに動詞として認識され、換言パターンの適用を間違えた。

6. おわりに

本稿では、形態素解析結果のみを用いて、パターンに基づく換言処理を試みた。より多く換言文を生成するために、 換言コーパスから換言現象を分けて換言パターンを獲得する方法を提案し、また、多種の換言パターンを適用する仕組みを設計した。換言実験により、約 10% の入力文が換言され、生成した換言文は約 88% の正解率が得られた。

本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである.

参考文献

[Kat00] 片岡明, 増山繁, 山本和英：動詞型連体修飾表現の "N_1のN_2" への言い換え, 自然言語処理, Vol. 7, No. 4, pp. 79–98 (2000).

[Xia00] XIA, F.: The Part-of-speech tagging guideline for the Penn Chinese Treebank (3.0), *available at http://www.ldc.upenn.edu/ctb* (2000).

[Yam01] 山本和英, 白井諭, 坂本仁, 張玉潔：Sandglass: 両言語換言機構を基軸とする音声翻訳, 言語処理学会第 7 回年次大会発表論文集, pp. 221–224 (2001).

[Zha99] 張黎, 佐藤晴彦：中国語表現文法 -28 ポイント, 東方書店 (1999).

[Zha01a] ZHANG, Y. and YAMAMOTO, K.: Analysis of Chinese Spoken Language for Automatic Paraphrasing, In *ICCPOL'2001*, pp. 290–293 (2001).

[Zha01b] ZHANG, Y., ZONG, C., YAMAMOTO, K., and SAKAMOTO, M.: Paraphrasing Utterances by Reordering Words Using Semi-Automatically Acquired Patterns, In *NLPRS'2001*, pp. 195–202 (2001).