Internal Use Only（非公開）

TR-SLT-0009

中国語処理のための表現変換の研究

Research on the Paraphrasing of Chinese Utterances

張玉潔
Zhang Yujie

2002年3月28日

概要

中日音声翻訳において、音声認識部から中国語の音声言語の多様な表現を受け取り、限定された発話表現に言い換え、変換部に渡すことを目的とする。次のニ種類の言い換えを行なうことを目標とする:(1)会話表現、口語表現などの検出と換言(2)構文構造の簡単化と意味情報の特定。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288 京都府相楽郡精華町光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

目次

1 はじめに

近年自由発話の音声翻訳システムが盛んに研究されている。音声認識技術と機械翻訳技術の進展により、分野限定の音声翻訳システムが盛んに研究されている。注目される一つの課題は自由発話の音声翻訳システムであり、システムの実現と自由発話の研究との二つの方向から検討されている。システムの実現において、書き言葉を原点とし、書き言葉と外れた発話現象を対処出来るように現在の技術を最大に利用する。自由発話の研究において大量のデータを収集し分析を行うことが重視されている。自由発話、あるいはある程度の自由度により発話された音声言語には書き言葉と外れたさまざまな現象が出てくる。たとえば、雑音による音声認識の誤り、会話による言い直し、語順倒置などの現象がある。書き言葉に向け設計された機械翻訳の仕組みをそのまま利用すると、システムの音声言語に対する頑健性を保つことはとても困難である。

そこで、我々は次のように考える。ある効果に達成するための発話はいろいろな表現があり得る。逆に、それらの多様の表現はある単一の発話表現に言い換えることが出来る。これにより音声翻訳システムを次の三つの部分に分ける: (1) 原言語内部の換言部：音声言語の多様な表現を相対的に限定された発話表現に言い換える。 (2) 変換部：限定された発話表現を原言語から目的言語へ写像する。 (3) 目的言語内部の換言部：限定された発話表現を音声言語の多様な表現に言い換える。

このように、音声言語の多様性をその言語の内部で換言処理で解決することにして、変換部には限定された表現を異なる言語に転換することにする。音声言語を翻訳するという複雑な問題を単言語の換言処理と言語対の間の変換処理との二つの問題に分ける。

翻訳の前処理に関する研究が昔からなされている。日英機械翻訳において、日本語長文の短文分割と主語の補完を行うという研究と構文構造の曖昧さを減少するための原文の自動書き換えにより訳文の品質が向上したと報告されている。また最近、換言処理は多く研究され、いくつかの技術手法が発表されている。

現在、換言処理に基づく中日音声翻訳システムを構築するSandglassというプロジェクトを進めている。以下では、その中の中国語の換言処理について議論する。

2　換言処理の目的

中日音声翻訳システムにおいて、中国語の換言処理は書き言葉と外れた音声言語現象を対処するために設けられる処理部である。換言処理は音声認識部から認識結果を受け取り、同じ発話効果（意味）を持つ書き言葉に言い換え、変換部に渡す。換言処理は次の三種類の言い換えを行なうことを目標とする。

2.1　音声認識誤りの訂正

音声認識技術の急速な発展にも関わらず、雑音の環境下で、また読み上げではなく自由発話に対し音声認識の誤りは現段階でなくすことはできないだろう。音声認識の誤りが存在すると想定し、換言処理では音声認識の誤りへの対処を行うことが賢明であろう。

2.2　会話表現、口語表現などの検出と換言

会話の性質また自由発話により、言い直し、語順倒置などの現象が現れる。書き言葉を対象とする変換部にとってはこのような発話現象は複雑過ぎてそれを処理できなくなる。そこで換言処理は発話の中の不要語を削除して、同じ意味の書き言葉の表現に言い換える。また言い換え切れない情報があれば、それを抽出し補足情報として添付する。

2.3　構文構造の簡単化と意味情報の特定

構文構造の複雑さと語彙の曖昧さを解決することは従来の機械翻訳システムの難点である。変換部のこの部分の負担を減らすことも換言処理の一つの目的とする。例えば、曖昧な文型表現を簡単化するとか、訳語選択の難易度を軽減させるために単語を入れ替えるなどの対処を行う。さらに、省略に対する補完、照合に対する情報の特定も行うべきであると考える。

3 換言処理対象の分析

換言処理はどんな言語現象を対象するか、換言処理をどう実現するかという問題に関して、中国語の発話現象に対し先行分析を行った。今回の分析は主に自由発話の現象を中心にした。

3.1 分析データ

分析データとしてはATRの旅行会話日中対訳コーパスとLDCのCallHomeの書き起こしコーパス[1]を利用した。分析の目的に応じて、発話の自由度により三種類のデータに分けた。

A類) 書き言葉のデータ: ATRの旅行会話日中対訳コーパスの中の、日本語の発話を中国語に翻訳したものからなる。変換部はA類の文型現象に対処できるという目標を立てている。これにより、A類のような文型と外れた言語現象を換言処理の対象とし、これをA類のような文型に言い変えることを換言処理の目的とする。音声認識部も今の段階ではA類のデータを用いている。

B類) 話し言葉のデータ: ATRの旅行会話日中対訳コーパスの中の、部分的な中国語の書き言葉に対し、三人の中国語のネイティブにより言い換えたものからなる。会話の背景が考慮されているので、語順倒置現象が含まれる。B類のデータは話し言葉の語順倒置現象を分析するときに利用する。

C類) 自由発話のデータ: LDCのCallHomeの書き起こしコーパスを利用した。これは電話による中国語のネイティブの間の無台本会話からなる。会話内容は主にアメリカあるいは中国国内での勉強、仕事と生活に関わる。話者たちは親友の関係であるから、発話は非常に自由だと思われる。一つの会話が5~10分程度の録音時間のデータで、コーパス全部で120会話がある。そのうち訓練データとしての80会話において、述べ語彙数は155,276である。CallHomeの書き起こしコーパスは言い淀み、言い直しなどの現象を含んでいる。ATRの旅行会話日中対訳コーパスに欠如したこれらの現象を補うために、CallHomeのコーパスを参考にする。C類のデータは言い淀み、言い直しなど自由発話の独特の現象を分析するときに利用する。

3.2 音声認識の誤り

分析のため、A類の書き言葉のデータを読み上げて音声認識を行った。その結果のうち237文の認識結果を分析し、次のような誤りがあることが分かった。

- 文字の挿入: 全部で37個所、発生場所は文末が一番多く、文頭が二番目で、文中が少なく一カ所だけである。

[1] http://www.ldc.upenn.edu/Catalog/LDC96T16.html

[例3.2 – 1]
[テキスト] 现在是在东京酒店房间号码是七五 一
[認識結果] 现在是在东京酒店房间号码是七五一 要我

- 文字の置換: 全部で 15 個所、そのうち数字の置換は 8 個所があった。数字が同音異文字に置換された場合が多かった。
  [例3.2 – 2]
  [テキスト] 您的维萨卡号码是四八八三五八零零四零八八 一 七 一 八对吗
  [認識結果] 您的维萨卡号码是四八八三五八零零四零八八的 要 七 要 八对吗

- 理解不能: 文字の挿入、置換、欠落のため理解できない結果は 8 文あった。

3.3 言い淀み、言い直し

発話という行動に関していろいろな観測面がある。例えば話者の方面から見ると発話時の心理状態、発話の目的などがあり、発話効果の方面からあるいは聞き者から見ると情報の伝達、発話権の制御などがあり、発話の生成物から見ると発声した文字列がある。このような考えに基づいて C 類の自由発話のデータを分析した。

(3.3.1) 言い淀み

言い淀みという概念は主に発話による伝達された情報から考えたものであろう。つまり新しい情報にあまり貢献していない文字列のことである。その文字列が実際の内容をもつかどうかにより、二種類がある。

- フィラー: 実際の内容をもたないが、発話権を持つという役割がある。
  [例3.3 – 1]
  现在因为国内的 那个那个那个 金融市场还没有到那一步嘛.
  (現在国内では あのーあのーあのー 金融市場はまだその段階になっていないから。)

- 繰り返し: フィラーとの相違はその文字が実際の内容を持つ点である。
  [例3.3 – 2]
  就 去，去 华尔街啊.
  (それでウォール街に 行って、行って。)

(3.3.2) 言い直し

話者が誤って言った直前の内容を訂正するための発話である。その効果として、構文が正しくなったり、意味がより適切になったり、話しやすくなることがあると考えられる。

[例3.3 – 3]

不象你们那个 很嗯， 比较 糟糕.

(あなたたちが持っているもののように それほど、 比較的に 悪くない。)

[例3.3 – 4]

对， 一月份，一， 一月底 回去的.

(はい、 一月、一、 一月末 に帰ったんです。)

3.4 口語表現

C類のデータの中で言い淀み、言い直し以外にも、文型が書き言葉と違った口語独特の表現が観察された。

[例3.4 – 1]

我现在 反正是 初步打算 是 改行啦.

この中の「反正是」(いずれにせよ) は挿入語と考えられ、次の二つの書き言葉の表現に言い換えることが出来る。

[例3.4 – 2]

反正我现在初步打算改行啦. (いずれにせよ私は今一応転業するつもりだ。)

[例3.4 – 3]

反正我现在的初步打算是改行啦. (いずれにせよ私の今の基本的な考えは転業だ。)

3.5 語順の倒置

分析用のデータはB類の話し言葉のデータを用いた。語順の入れ替えに着目して、書き言葉の文と対応する話し言葉の文を分析し、話し言葉ではいくつかの語順倒置現象があることが分かった。以下の例はその中の一部分である。

(1) 前置詞句

[例3.5 – 1]

[書き言葉] 对房间 有什么要求吗？ (お部屋の 御希望はございますか)

[話し言葉] 有什么要求吗？ 对房间

(2) 助動詞句

[例3.5 – 2]

[書き言葉] 我应该 告诉谁好呢？ (誰に知らせ ればいい ですか)

[話し言葉] 告诉谁好呢？ 我应该

(3) 条件節

[例3.5 – 3]

[書き言葉] 好的，那么，要是来晚的话，请再打电话给我好吗？
(かしこまりましたではもし遅れられるようでしたらまたお電話いただけますでしょうか)
[話し言葉] 好的，那么，请再打电话给我好吗？要是来晚的话.

3.6 日本語発話との比較

中国語文の基本的な語順は英語と同様SVOであるが、連体節とそれが修飾する体言との前後関係は日本語と同じである。実際特に話し言葉では、[例11]の示されるように語順が大体日本語の語順と同じような発話があり得る。

[例3.6 − 1]
[日本語] あっ、はい。パン屋は800円です、時給。
[中国語] 啊，对，面包房 (パン屋) 800 块钱 (円)，小时工资 (時給).

日本語の言い淀みの語が接続詞「で」と共起した現象に着目した研究が報告されている。その報告に引用された例を見ると言い淀みが文節の境界でしか現れない。中国語の場合では、C類の自由発話のデータにおいて言い淀みが単語の境界で現れていた。

言い直しに関しては、日本語の発話において文中の任意位置で言い直しが現れる。一方中国語の発話においては、単語の途中からの言い直しが観察され、言い直しは文中の任意位置で現れることが可能である。

3.7 換言研究の課題

換言処理実現のために我々が今後どのような検討を行う必要があるかを考察する。

- 換言処理では、単語レベルでの不要語の削除、多義語の入れ替えがあり、口語表現による構文レベルでの文の生成もある。いずれも話し言葉から書き言葉への換言知識が必要である。そこで話し言葉のさまざまな表現に対しデータの収集と分類が必要になる。

- 換言ルールの作成と管理を自動化するために、書き言葉の文と対応する話し言葉の文の対が大量に必要である。現在ATRでは人手による書き言葉から話し言葉への言い換え作業を行っている。すでに二万文から四万文の言い換え文が得られた。その以外に、換言技術を利用して話し言葉の文を自動的に生成させ、人手の後修正を加えるという手段も考えられる。

- 音声翻訳における原言語の換言処理が音声認識部と変換部と関係するので、両方の当面の問題を対処させながら換言部を開発していくべきと考える。例えば、現在の認識結果では数字の置換があるので、このような誤りを対処できるような換言処理を実現する必要がある。

- 換言処理では同じ意味の文を生成するために構文解析が必要である。無論必要最小限の解析のみを行うのが理想である。日本語と違い、入手可能な中国語解析ツールは (中国を含め) ほとんどない。そこで浅い (shallow, partial) 構文解析ツールの開発が必要である。

4 換言処理の手法

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よう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、換言処理は全文の意味を理解しなくてもできる。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。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。

5 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。このコーパスは換言に関する貴重な言語資源で、換言知識の獲得には充分に利用すべきである。換言コーパスの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。このレポートで現れた品詞の部分を表1に示す。

表 1. Part of the part-of-speech set of the Penn Chinese Treebank

Symbol	Explanation
NN	common noun
NR	proper noun
NT	temporal noun
PN	pronoun
DT	determiner
DEC	的 in a relative-clause
DEG	associative 的
M	measure word
JJ	other noun-modifier
VA	predicative adjective
VC	是
VE	有 as the main verb
VV	other verb
AD	adverb
P	preposition excl. 被 and 把
LC	localizer
CD	cardinal number
OD	ordinal number
SP	sentence-final particle
BA	把 in ba-construction
CC	coordinating conjunction

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。

5.1 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。

(5.1.1) 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。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5.1 − 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

(5.1.2) 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5.1 − 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)

(5.1.3) 文法標識“把”

中国語では、日本語のように、“が”は主語を、”を”は目的語を指示するような”表層的な文法文法マークがない。前置詞“把”はその一つである。“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ＯＶC”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5.1－3]

[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)

5.2 換言文対の自動抽象化

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5.1-3の換言文対を抽象化して例5.2-1のパターンが得られる。

[例5.2－1]

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。

5.3 換言文対の半自動抽象化

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5.3－1のように換言文対が定義され、例5.3－2のパターンが得られる。

[例5.3－1]

[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5.3－2]

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 5.2-1 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。

5.4 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。

6 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。

7 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2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7－1]
[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7－2]
[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7-1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7-2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

表2. 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%)

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。

8 おわりに

換言処理は新しい研究課題で、音声翻訳ための中国語の換言処理は始めたばかりである。本稿では、中国語の換言処理の試みを紹介した。より多く換言文を生成することを目指して、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。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た。より正確に、抽象度が高いパターンを得るために、換言文対から換言パターンを半自動化に獲得するツールを開発した。開発したツールを用いて、約2800個の換言パターンを得た。また、多種の換言パターンを適用する換言装置の仕組みを設計し、実現した。45100の評価文における換言実験の結果により、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。この換言手法では、深い解析と複雑な文の生成を避けることが出来て、処理時間が速い。開発したパターン作成ツールは換言文対からパターンを獲得するには人の知識を組み込むことが出来て、カバーレッジがより高いかつより正しい換言パターンを得ることができた。