中日音声翻訳のための中国語換言処理の分析

張 玉潔　山本 和英　坂本 仁
E-mail: {yzhang, yamamoto, msakamo}@slt.atr.co.jp
ATR音声言語通信研究所

1　はじめに

本稿では換言処理(paraphrasing)に重点を置いた音声翻訳手法を提案し、中国語の換言処理について考察する。

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。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、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[Shi95]。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[Sat99]。

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[Yam01]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。また、中国語の発話現象を分析して、今後の研究について考察する。

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。

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

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[Zha00]。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。
  [例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] 您的维萨卡号码是四八八三五八零零四零八八二七二八对吗
  [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] 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4] 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5] 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6] 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7.1] 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7.2] 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7.3] 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8]
[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9]
[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10]
[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.

4 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11]
[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 [Tsu99]。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 [Ara99]。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。

5 今後の研究についての考察

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。

6 おわりに

本稿では、換言処理に重点をおいた音声翻訳手法を提案した。またその応用としての中日音声翻訳の実現に向けて、中国語の発話現象を調査し、中国語の換言処理の対象に関して検討した。さらに今後の換言処理の研究方向について考察した。

参考文献

[Ara99] 荒木哲郎, 池原悟, 三品尚登：N-gram を用いた対話文の言い直し表現の検出法, 自然言語処理, Vol. 6, No. 3, pp. 23–41 (1999).

[Sat99] 佐藤理史：論文表題を言い換える, 情報処理学会論文誌, Vol. 40, No. 7, pp. 2937–2945 (1999).

[Shi95] 白井諭, 池原悟, 河岡司, 中村行宏：日英機械翻訳における原文自動書き替え型翻訳方式とその効果, 情報処理学会論文誌, Vol. 36, No. 1, pp. 12–21 (1995).

[Tsu99] 土屋菜穂子：対話コーパスを用いた言い淀みの語の統語論的考察, 言語処理学会第 5 回年次大会発表論文集, pp. 251–254 (1999).

[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, A4-1 (2001).

[Zha00] ZHANG, S., ZHANG, J., NAKAMURA, S., and SAGISAKA, Y.: A Preliminary Investigation of Sub-Syllabic Modeling for Chinese Speech Recognition Based on HM-NET, 日本音響学会 2000 年秋季研究発表会講演論文集, pp. 127–128 (2000).