# 用例に基づく換言：中日旅行会話翻訳への適用

大竹 清敬
ATR 音声言語コミュニケーション研究所
kiyonori.ohtake@atr.co.jp

## 1 はじめに

現在，我々は中日旅行会話音声翻訳機に関する研究開発を進めている．この翻訳機の特徴は，翻訳の頑健性を向上させるために換言を用いているところにある [Yam02]．この翻訳機では，換言が，原言語側と目的言語側でそれぞれ行われる．換言の目的は，原言語側では，言語変換器が翻訳しやすいように，そして目的言語側では翻訳の場面，状況に適した表現となるように表現を変換することである．

音声翻訳においては，音声認識器の認識誤りをはじめとした様々な誤りに対して頑健であることが強く望まれる．そのため，表現の変換を行う換言処理もこれらの誤りに対して頑健でなければならない．したがって，不適格あるいは不自然な表現を適格かつ自然な表現へ変換する技術が要求される．

換言は同一言語内での翻訳ととらえることができる．そのため，これまで機械翻訳を実現するために考えられてきた様々な手法を利用可能である．現在の主要な機械翻訳手法として，規則に基づく方法，用例に基づく方法，統計的な方法などがある．どのような誤りが存在するか不明な対象に対して多様な換言を実現しようとしたとき，規則に基づく換言では，規則の収集があまりにも高価に，そして煩雑になる．そこで，我々は翻訳プロセス全体を確実に制御したいという要求から用例に基づく換言について検討した．

中日音声翻訳の目的言語側における換言の目的はすでに述べた通りである．しかしながら，現実には翻訳機が完璧な訳を出力することは考えられず，不自然，あるいは不適格な出力をすると考える方が自然である．そこで，この報告では，用例に基づく換言を用いて中日音声翻訳の目的言語側の言語表現をより自然な表現へと変換する手法について述べる．

この報告で報告する日本語換言器は，別の言い方をすると，被換言表現が使用される状況がある程度限定された条件のもとで，もっともらしい表現へ換言するものとも言える．その基本的な動作は，被換言表現に類似する用例をデータベースから探し，適用することである．

用例に基づく換言を適用することによって，音声翻訳に関するいくつかの利点が生じる．その一つは，コーパスを用いる利点を最大限に活かせることである．換言処理は単言語処理であるので，コーパスを収集しやすい．また，換言を適用する場面ごとに用例をあらかじめ分類しておくことによって，適用すべき用例の選択に対して自然な制約を実現できる．すでに，このような「場面」に相当するクラスを発話単位で推定する手法は提案されており [Asa02]，換言の精度を向上させるために非常に有用である．

さらに，翻訳の後処理という観点から換言を捉えるならば，本報告で提案する換言処理は，翻訳機や相手言語に依存しない．したがって，日本語を目的言語とするあらゆる翻訳機の出力に適用することが可能である．

## 2 用例に基づく換言

用例に基づく換言の基本的な枠組みは用例に基づく機械翻訳と同一である．換言器を機能させるためには次の3つの処理が必要である．

**収集** 目標とすべき表現を収集する．
**検索** 被換言表現に最も類似した表現をデータベースから検索する．
**適用** 被換言表現と得られた用例との違いを考慮し，用例を適用する．

以下それぞれの処理について説明する．

[収集]

用例の収集の際には，各用例に対して2つの処理を行う．汎化と換言である．用例には数詞や，固有名詞など，各用例を特徴づける単語が含まれるが，用例の活用という点からは重要度が低いため，それらの単語を汎化する．汎化は，数詞，固有名詞（人名地名など），日時を示す表現（月，曜日）などを形態素解析結果を用いて記号列へ置換することにより行う．また，ここで適用する換言は，主に用言に関する表現の多様性を確保する目的で行う．現在，文献 [Oht01] にて示した手法を用いているが，今後充実させる．

[検索]

音声翻訳処理における翻訳結果に対して換言を適用しようとする状況では，翻訳機，あるいはその前段の処理における誤りによって被換言表現を正しく解析できない可能性が存在する．そこで，用例を検索する場合には被換言表現に対して形態素解析などを行わず，文字列を検索する．検索単位は，被換言表現から n-gram 文

字列を抽出し，それを検索する．現在，経験的に n=3 が良いことがわかっている．たとえば，「ホテルの予約」からは「ホテル，テルの，ルの予，の予約」が検索文字列として抽出される．用例はすべて汎化されているため，被検索文字列も同様に汎化しておく．

さて，汎化した被検索文字列を $S$ とし，$S$ に含まれる検索単位の文字列の集合を $T(S)$ と表記する．すべての $t \in T(S)$ についてデータベースを検索する．検索文字列 $t$ による用例検索結果の集合を $R(t)$ とする．すべての $t \in T(S)$ について $E_i \in R(t)$ である用例 $E_i$ に関して

$$W(E_i) = d_{len}(|S|, |E_i|) \sum_{t \in T(S) \cap T(E_i)} \frac{1}{\log(|R(t)|+1)+1} \quad (1)$$

を求める．$W(E_i)$ を用例 $E_i$ のスコアとする．関数 $d_{len}$ は長さに関する制約を適用するためのものであり，次式で定義される．

$$d_{len}(a,b) = \begin{cases} a/b & if\ a < b, \\ b/a & otherwise. \end{cases} \quad (2)$$

この $W(E)$ の大きい順に上位 $M$ 用例について，$S$ との編集距離 $ED(S, E_i)$ を求める．$ED(S, E_i)$ は文字を単位とする距離である．編集距離を求める理由は，n-gram 文字列による検索のみでは構成する文字列が類似する用例を検索できても，その並びが被換言表現に近い保証がないためである．最終的には，$M$ 用例のうち $ED(S, E_i)/(|S| + |E_i|)$ が最も大きい用例を適用候補として採用する．

なお，検索結果が得られなかった場合，または最終的なスコアが既定の閾値に満たない場合は，換言処理を放棄する．

[適用]

選択された用例には，汎化した記号列が含まれるためこれを被換言表現に基づき復元する．汎化記号列の復元に失敗した場合は換言処理を放棄する．さらに，被換言表現と採用された用例との間にある不一致箇所について詳細に検証する．特に，汎化されない名詞については階層的シソーラスなどを導入し，置換する必要がある．たとえば，被換言表現「トリプルルーム空きあるか」に対して用例が「和室は空いてますか」だった場合に，「トリプルルーム」⇔「和室」を置換しなければならない．残念ながら，この置換処理は現在未実装である．

# 3 実験

音声翻訳における用例に基づく換言処理の有用性を確認するために小規模ながら実験を行った．

## 3.1 中日翻訳機

ここでは，実験に使用した中日翻訳機の概要について簡単に説明する．この翻訳機は，中日の対訳コーパスから翻訳知識（翻訳パターン）を自動獲得し，それを用いることによって翻訳を行う．基本的な考え方は，文献 [McT03] と同様である．必要とする言語資源は中日の対訳コーパスと中日の辞書である．また，翻訳パターンを獲得する際に日本語形態素解析器を必要とする．

中日翻訳機は，入力された中国語発話に対して，順次適用できる翻訳パターンを適用し，日本語訳を生成する．翻訳パターンにおいて変数として扱われる単語に対しては中日辞書を参照し，日本語訳を決定する．ただし，このとき，複数の日本語訳が存在する場合でも，翻訳機は訳語選択を行わずそのまま全ての訳語を埋め込んだ翻訳結果を出力する．

## 3.2 コーパス

用例データベース作成のために用いたコーパスは ATR 音声データベース[1] に含まれる約 74000 発話，ならびに ATR 旅行会話基本表現集 (BTEC)[Tak02] に含まれる約 108000 発話である．この 182000 件の用例に対して，換言 [Oht01] ならびに汎化を適用した結果，約 337000 用例となった．文献 [Oht01] に示した換言を行わず，汎化のみの場合は 174000 用例となる．

中日翻訳が翻訳パターンを抽出するために使用したコーパスは，ATR 音声データベースに含まれる日本語発話を人手で中国語翻訳して作成した対訳コーパスの中から抽出した 2958 対訳である．この対訳コーパスから得られた翻訳パターンは約 35600 である．なお，対訳コーパスに含まれる日本語発話のすべては換言器の用例データベースに含まれる．

## 3.3 諸条件と実験結果

実験は，翻訳機が学習に用いたコーパスに含まれる全ての中国語発話を翻訳機を用いて翻訳させ，それを換言器を用いて換言した結果を評価する．まず，中日翻訳機による翻訳結果は訳語選択が行われていないため，前後の文字列を用いる簡便な訳語選択手法を適用し，訳文を完成させる．この訳文とそれに対応する元の正解となる日本語発話との文字単位の編集距離が 7 以上である 334 発話から 100 発話を任意に選択し，これを換言する．これら 100 発話の翻訳結果と換言結果を，比較評価する．結果を表 1 に示す．また，これら 100 発話の翻訳結果のうち翻訳の修正あるいは換言による回復を要求する不適格，不自然な結果は 53 件であった．換言事例を表 2 に示す．

今回の実験の設定では，翻訳結果が対応する正解文

[1] http://www.red.atr.co.jp/detabase.html

表 2: 換言事例

| | | |
|---|---|---|
| 無効果 | T | ええかしこまりましたカード番号をお願いできますでしょうか. |
| | P | はい, かしこまりました. カードのナンバーをお願いできますでようか. |
| 改善 | T | カード四八八三五八零零四零八八一七一八はビザです. |
| | P | カードはビザで, 四八八三五八零零四零八八一七一八です. |
| | T | あす妻一緒参加京都にのバス観光ツアーにしたいんです. |
| | P | あす, 妻と一緒に, 京都のバス観光ツアーに参加したいんです. |
| 改悪 | T | それからサービス料はついてるんですか. |
| | P | それから朝食はついてるんですか. |

T: 翻訳結果　　P: 換言結果

表 1: 評価結果

| | 換言により変化 | 改善例 | 改悪例 |
|---|---|---|---|
| 事前換言あり | 88 | 36 | 16 |
| 事前換言なし | 89 | 36 | 18 |

は換言器の用例データベースに含まれるため事前換言による大きな違いはなかった. 以降では, 事前換言ありの場合について詳細に検討する.

まず, 換言による変化がなかった事例が12件あるが, 換言できなった理由は次の2つである. (1) 翻訳結果に問題はなく, データベース中の用例と完全に一致した:4件. (2) 翻訳結果に対する適当な用例を見つけることができず, 換言を放棄した:8件.

つぎに, 換言した結果, 改悪となった事例16件について調査した. これらの事例は次の5つの理由が主な原因だった.

(a) 選択した用例は良いが, まだ実装していない名詞の置換を必要とする:2件. (例) もしできましたらトリプルルームをお願いいたしたいんですが。→ もしできましたら和室をお願いいたしたいんですが。

(b) 翻訳結果が悪く[2], 適当な用例を選択できなかった:6件. (例) ところでたいが十一日と十二号それで結構ですですかをお願いします。→ じゃあ、それで結構ですので、そちらでお願いします。

(c) 表層的な違い (たとえば,「今日」⇔「きょう」) によって適当な用例を選択できなかった:2件. (例) わたしのルームナンバーは五百十五名前はエイミーハリスです。→ 私のルームナンバーは何番ですか。(該当用例:わたしの部屋番号はNです。名前はPです[3]。)

(d) 翻訳結果が不十分[4]であるため適当な用例を選択できなかった:4件. (例) お診察下さい。→お待ち下さい。(正解訳:診察をお願いしたんですが。)

(e) 訳語選択に失敗し, その結果適当な用例を選択できなかった:2件. (例) 求めはできますか。→返品はできますか。(正解訳:リクエストはできますか。, 訳語候補:{ リクエスト, リクエストする, 依頼, 求め, 需要, 請求する, 要求, 要求する, 要望, 要望する })

[2]翻訳結果が悪いとは, その発話単体では人間にも伝えたい内容がよくわからないという意味である.

[3]この例では, 数詞がNに, 固有名詞がPに汎化されている.

[4]翻訳結果が不十分とは, 伝えたい内容は人間にはわかるが通常用いない表現をしているという意味である.

## 4 考察

まず, 翻訳機も含めて全てが開発途上であることを考えると, この実験結果は楽観していいものであると考える. その理由は, 換言によって改悪された事例が16件と相対的に多いものの, これらの多くは翻訳の質が低い事に起因するからである. また, 改悪された事例の多くは, 非常に長い発話である. 現段階では処理単位を明確に設定していないため, このような長い発話も処理している. しかし, 最終的には, 規定された処理単位において, 原言語側の入力が分割可能な場合には原言語側の換言器によって文分割を行う. この結果, 翻訳機の負荷が軽減され, 翻訳の質も向上すると予想する. 翻訳機単体での翻訳品質の改善も予定されていることから, 翻訳の質に起因する改悪事例は減少すると考える.

つぎに, 換言によって変化が生じた事例88件のうち約4割に当たる36件が特に効果が認められない換言であった. これは, 実験では, 日本語旅行会話とそれを中国語に翻訳した対訳コーパスを用いて中日翻訳機を構成し, 元の日本語旅行会話コーパスは換言器の用例として利用されたため, 当然の結果といえる. 逆に, 現在の中日翻訳機が出力する日本語のうち, 正解と7以上の編集距離を持つ出力100件中に不適格, 不自然な訳が53件あり, そのうちの36件を換言によって適格かつ自然な発話に変換できた.

しかしながら, 以上の結果は, 翻訳結果の正解訳となる表現が換言器の用例データベース内にすべて存在するという条件下での結果である. 今後は, 翻訳機も含めて未知の表現が入力された場合に翻訳機ならびに換言器がどのように振舞い, そしてどのような問題が起こりうるかを詳細に調査する必要がある.

換言器における大きな問題のひとつは, 入力とそれに適用する用例との間にある不一致表現にどのように

対処するかである．現状では，汎化されない一般名詞の違いへの対処は必須である．また入力と用例との間に表現の過不足がある場合，入力，用例のどちらがその状況に適しているかを判断して調整しなければならない．このような高度な対処を行うためには，入力を解析する方が有利ではあるが，現在有している頑健性を保持することが困難になる．

一方で，換言器が用いる用例は比較的収集しやすいものの，用例として用いるだけの十分な品質を保証する技術は確立されてはいない．そのため，収集したコーパスを検証し，品質を保証する必要がある．これは，コーパスのみならず，コーパスに対して適用する事前換言に関しても同様である．さらに，この問題はコーパスに基づいて知識を獲得する翻訳機にも存在し，検証が必要なのは同様である．しかしながら，処理対象が単一言語である換言器のコーパスに対して，翻訳機が対象とするのは対訳コーパスであるため検証が高価になることが予想される．

## 5 関連研究と応用

用例に基づく換言手法の基本的な枠組みは用例に基づく翻訳手法と同じである．しかしながら，事前換言によって処理する言語表現の多様性を確保しようとしている点が特徴である．この結果，この換言器を適用可能な自然言語処理では，さまざまな言語表現に対応可能となり頑健性が向上する．この報告では，事前換言に適用可能な換言技術について詳細に述べていないが，換言に関する詳しい研究動向は文献 [Inu02] を参照されたい．

また，用例に基づく換言技術は様々な応用が可能である．たとえば，用例を用いて音声認識における誤り訂正を行う研究として沖本ら [Oki01] を挙げる．本報告では，用例に基づく換言を中日翻訳の出力に適用し，目的言語側の表現を適格かつ自然な表現へと変換することを試みた．逆に原言語側の換言は，翻訳機が翻訳可能な言語をある制限言語とみなすと，図 1 に示すように実現可能である．入力表現が制限言語に含まれない場合に，それらを換言したコーパスを検索し，該当表現のリンクをたどり，元となった制限言語を用いる．制限言語と換言コーパスの間の関係が多対多となる換言を適用できれば，柔軟性に富んだ，頑健性の高い換言が実現できると考える．

## 6 まとめ

本報告では，用例に基づく一換言手法を示し，それを中日旅行会話翻訳の出力に適用する実験を行った．正解訳との編集距離が 7 以上の翻訳結果 334 発話から 100 発話を任意に選択し，換言を適用した．その結果，翻訳誤りなどにより修正を必要とする 53 発話のうち 36 発話を換言によって適格かつ自然な表現へと変換できた．残りの 17 発話の多くは翻訳誤りが原因で正しく換言できなかった．

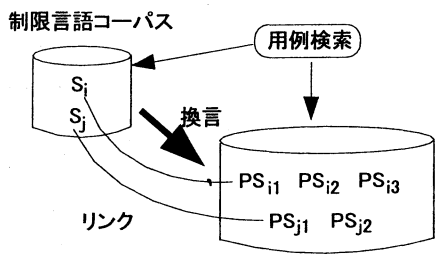


図 1: 用例に基づく制限言語への換言

実験結果は，換言器にとって楽観できるものと考えるが，翻訳の後処理として換言を適用するためには解決しなければならない問題も残されている．具体的な今後の課題としては，シソーラスなどを用いた一般名詞の置換，一般名詞以外の不一致箇所の調節がある．

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

[Asa02] 浅見克志, 竹澤寿幸, 菊井玄一郎：音声対話処理のための発話単位のトピック推定, 情報処理学会研究報告 SLP-42 (2002).

[Inu02] 乾健太郎：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).

[McT03] McTait, K.: Translation Patterns, Linguistic Knowledge and Complexity in an Approach to EBMT, In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pp. 299–329, Kluwer Academic Press (2003), (forthcoming).

[Oht01] Ohtake, K. and Yamamoto, K.: Paraphrasing Honorifics, In *Workshop Proceedings of Automatic Paraphrasing: Theories and Applications (NLPRS2001 Post-Conference Workshop)*, pp. 13–20 (2001).

[Oki01] 沖本純幸, 山本博史, 隅田英一郎, 菊井玄一郎：旅行会話基本表現コーパスを用いた認識誤り訂正の検討, 電子情報通信学会 信学技報 NLC2001-12, pp. 49–54 (2001).

[Tak02] Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S.: Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Converstaions in the Real World, In *Proceedings of LREC 2002* (2002).

[Yam02] Yamamoto, K.: Machine Translation by Intraction between Paraphraser and Transfer, In *Proceedings of COLING2002*, pp. 1107–1113 (2002).