翻訳リペア支援のための言い換え文自動生成手法の提案

宮部　真衣†　　吉野　孝‡,††
†和歌山大学大学院システム工学研究科　　‡和歌山大学システム工学部
††（独）情報通信研究機構言語グリッドプロジェクト

1　はじめに

世界規模のインターネットの普及により，ネットワークを介した多言語コミュニケーションの機会が増加している．しかし，一般に多言語を十分に習得することは容易ではない．母語を用いた多言語間コミュニケーションを実現するために，機械翻訳を利用した取り組みが現在行われている．近年，機械翻訳技術は急速に進展しているが，完璧な翻訳を行うことは困難である．翻訳文中の不適切な翻訳箇所を減少させるために，入力文章を書き換えていくことを「翻訳リペア」と呼ぶ．

折り返し翻訳を用いた翻訳リペアにより得られる翻訳結果の精度検証実験から，翻訳リペアによって翻訳精度が改善できることが確認できている [1]．一方，翻訳リペア作業はユーザにとって負担のかかる作業である [2]．翻訳の不適切な箇所の強調表示による翻訳リペア支援が行われているが [3, 4]，翻訳の不適切な個所を強調したとしても，ユーザは修正内容を自分で考え出さなければならない．しかし，どのように修正するかを考え出すことはユーザにとって負担の大きい作業であり，修正作業の支援が必要とされている [4]．

そこで本稿では，ユーザの翻訳リペア作業を支援するために，翻訳精度の向上する可能性のある言い換え文を提供するための仕組みを提案する．

2　翻訳リペア支援のための言い換え文自動生成

翻訳リペアにおける主な作業は，動詞や名詞などを同義の表現へと変更する作業である [1]．そこで，類語辞書から単語の類語および関連語（本稿では，「言い換え候補」と呼ぶ）を取得し，その言い換え候補を用いて言い換え文を生成する仕組みを構築する．本研究では，形態素解析器，類語辞書，Web 日本語 N グラム [5]，機械翻訳システムを連携し，言い換え文の生成を実現する．言い換え文生成の流れを図 1 に示す．「翻訳不適箇所の推定」「言い換え候補の抽出」「言い換え文の生成」という 3 つの手順により言い換え文を生成する．

2.1　翻訳不適箇所の推定

入力文中の単語のうち，折り返し翻訳文中に存在せず，その言い換え候補も折り返し翻訳文中に存在しない単語を，言い換えの必要な翻訳不適箇所と判断する．

翻訳不適箇所の推定手順を以下に示す．

a. 入力文の折り返し翻訳文を取得する．
b. 形態素解析器 MeCab[6] を利用し，入力文および折り返し翻訳文の形態素解析を行う．
c. 入力文中の単語のうち，折り返し翻訳文中に存在しない単語を抽出する．

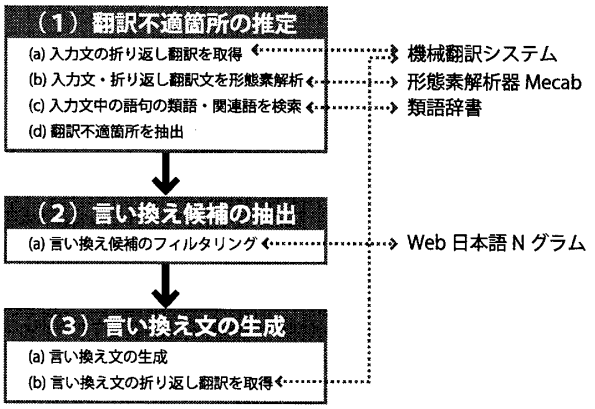

図 1: 言い換え文生成の流れ

d. 抽出した単語の言い換え候補をインターネット上の類語辞書[*1]により検索する．
e. 取得した言い換え候補が折り返し翻訳文中に存在しない場合，その言い換え候補のもととなる単語を翻訳不適箇所とする．

2.2　言い換え候補の抽出

言い換え候補の数はもとになる語によって異なるため，多数の言い換え候補が取得される可能性がある．また，もとの単語と言い換え候補を置き換えた場合，不自然な文になる場合もある．そのため，できる限り不要な候補を取り除く必要がある．

そこで，不適切な言い換え候補を除外するために，Web 日本語 N グラム [5] を用いる．Web 日本語 N グラムは，言語資源協会が発行している，日本語の単語 n-gram とその出現頻度をまとめた大規模言語リソース[*2]である．今回は，言い換え対象語およびその前後の品詞の組み合わせに関する出現頻度を調べるために，3-gram のデータ（異なり 3-gram 数は約 3.9 億）を用いる．また，翻訳不適箇所が文頭の場合については，3-gram のデータが利用できないため，2-gram のデータ（異なり 2-gram 数は約 8 千万）を用いてフィルタリングを行う．

言い換え候補の抽出手順を以下に示す．

a. 入力文中の翻訳不適箇所とその言い換え候補を置き換える．
b. Web 日本語 N グラムを利用し，置き換えた言い換え候補とその前後の単語の組み合わせの出現頻度を求める．
c. 出現頻度が 0 である場合，不要な候補であると判断し，言い換え候補から外す．

2.3　言い換え文の生成

抽出した言い換え候補をもとに，言い換え文の生成を行う．言い換え文生成の手順を以下に示す．

Automatic Generation Method of Paraphrased Sentence for Translation Repair Support
Mai MIYABE† Takashi YOSHINO‡,††
†Graduate School of Systems Engineering, Wakayama University
‡Faculty of Systems Engineering, Wakayama University
††Language Grid Project, National Institute of Information and Communications Technology

[*1] Yahoo!辞書を利用
[*2] n-gram データは，2007 年 7 月の Web ページのスナップショットから構築されており，用いられた総文数は約 200 億文である．このデータは出現頻度が 20 回以上の n-gram を抽出対象にしている．

表 1: 実験に用いたテキストの一部

機械翻訳試験文	・首相は経済をめぐる諸問題について語った。 ・ここで私は体積の変化は考えないものとする。 ・朝から降り続いた雨は、夜になって止んだ。
会話表現データベース	・チェックアウトする時の請求書を今用意してください。 ・クルージングの申込みはここでできますか。 ・今夜のお勧め料理はなにか教えてください。

表 2: 言い換え文により翻訳精度が向上した入力文数

テキスト	言い換え文の生成された入力文数（文）	翻訳精度が向上した入力文数（文）*
機械翻訳試験文	161	35
会話表現データベース	144	37

* 言い換えを行っていない入力文に対する折り返し翻訳文と比較して，言い換え文の折り返し翻訳文における一部分のみでも精度の改善が見られた場合は，翻訳精度が向上したと判断している．

表 3: 精度が向上した言い換え文の例

	言い換え前	言い換え後
入力文	現地の人々の気持ちも踏まえて行動したいものだ。	現地の人々の心境も踏まえて行動したいものだ。
折り返し翻訳文	それはまた人々のローカルな感覚に基づき、私は、ふるまいたい。	それはまた地元の人々の気持ちに基づき、私は、ふるまいたい。

言い換えられた単語は，各入力文における下線部の部分である．

a. 入力文中の翻訳不適箇所とその言い換え候補を置き換える．
b. 言い換え文の折り返し翻訳文を取得する．

3 言い換え文の生成実験

生成した言い換え文による翻訳精度の改善効果を検証するために，言い換え文生成実験を行った．折り返し翻訳文を取得するための機械翻訳システムとして，言語グリッド [7] を介して高電社の J-Server[8] を用いた．

翻訳するテキストとして，機械翻訳試験文 [9] および会話表現データベース [10] のうち，20 文字から 30 文字のテキストをそれぞれ 200 文を用いた．機械翻訳試験文については，該当文字数であるテキストを上から 200 文選択した．会話表現データベースについては，全トピックのテキストの順序をランダムにした後，該当文字数であるテキストを上から 200 文選択した．利用したテキストの一部を表 1 に示す．テキスト 400 文のうち，言い換え候補が存在した入力文は 374 文（機械翻訳試験文は 191 文，会話表現データベースは 183 文）であった．374 文において，言い換え候補の総数（重複する候補を含む）は 5782 語（機械翻訳試験文は 2964 語，会話表現データベースは 2818 語）であった．

表 2 に言い換え文の生成された入力文数と，言い換え文により精度の向上が見られた入力文数を示す．また，表 3 に精度が向上した言い換え文の例を示す．実験の結果，機械翻訳試験文は 34 文（17.6%），会話表現データベースは 37 文（20.2%）の入力文に対して，精度が向上可能な言い換え文が生成できた．また，言い換え文によって精度向上が見られなかった入力文について確認を行ったところ，以下の傾向が見られた．

(1) 一部の単語の翻訳ミスにより，折り返し翻訳文が低精度になっているが，得られた言い換え候補ではその翻訳ミスを改善できない．
例 入力文：新潮社のカセットブックが火付け役になった。
折り返し翻訳文：新潮社のオーディオブックはたいまつであった。

表 4: 言い換え文により翻訳精度が向上しなかった入力文に対する折り返し翻訳文の精度低下の原因

テキスト	言い換えにより精度が向上した文（文）	言い換えにより翻訳精度が向上しなかった文 *		合計（文）
		原因 (1)（文）	原因 (2)（文）	
機械翻訳試験文	35	125	40	200
会話表現データベース	37	96	67	200

* 言い換えにより翻訳精度が向上しなかった文は，以下の 2 種類の入力文を含む．
- 言い換え候補が存在せず，言い換え文が生成されていない文
- 言い換え文が生成されたが，翻訳精度が向上しなかった文

原因 (1)：入力文中の単語あるいはその同義の表現が折り返し翻訳文中に存在していない．
原因 (2)：入力文中の単語あるいはその同義の表現が折り返し翻訳文中に存在しているが，折り返し翻訳文の構造がおかしくなっている．

(2) 各単語は正しく翻訳されているが，折り返し翻訳文における文の構造がおかしいため，単純な単語の言い換えのみでは精度の改善ができない．
例 入力文：空港からホテルまでの所要時間はどのぐらいですか。
折り返し翻訳文：どれくらい必要とされている時間は空港からホテルまでであるか？

また，上記の原因に該当する，精度向上が見られなかった入力文の数を表 4 に示す．(1) については，言い換え候補の不足を補うことにより，改善できる可能性がある．今後，複数の類語辞書を用いて，単語の言い換えができるようにする必要がある．一方，(2) については，今回提案した仕組みで対応することは困難である．今後，構文解析等を利用し，単語の言い換え以外の言い換えに対応する必要がある．

4 おわりに

本稿では，修正の必要な単語の類語や関連語により，自動的に言い換え文を生成する手法を提案した．提案手法を用いて実験を行った結果，17〜20%の入力文に対して，精度の向上が可能な言い換え文が生成できた．

今後は，言い換え候補の不足に対する対応や，入力文の構造に問題のある場合への対応を行い，より精度の高い言い換え文の生成手法の検討を進める．

謝辞

本研究の一部は，独立行政法人科学技術振興機構「平成２１年度シーズ発掘試験Ａ（発掘型）」の補助を受けた．

参考文献

[1] 宮部真衣 ほか：折返し翻訳を用いた翻訳リペアの効果，信学論，Vol.J-90-D-I，No.12，pp.3142-3150（2007）．
[2] 小倉健太郎 ほか：目的指向の異言語間コミュニケーションにおける機械翻訳の有効性の分析：異文化コラボレーション ICE2002 実証実験から，第 65 回情報処理学会全国大会論文集，第 5 分冊，pp.315-318（2003）．
[3] 林田尚子 ほか：翻訳エージェントによる自己主導型リペア支援の性能予測，信学論，Vol.J88-D-I，No.9，pp.1459-1466（2005）．
[4] 宮部真衣 ほか：翻訳不適箇所の指摘による翻訳リペア効率の改善効果の検証，情処論，Vol.50，No. 4，pp.1390-1398（2009）．
[5] 工藤拓 ほか：Web 日本語 N グラム第 1 版，言語資源協会発行（2007）．
[6] Taku Kudo, et al.: Applying Conditional Random Fields to Japanese Morphological Analysis, EMNLP-2004, pp.230-237(2004).
[7] Toru Ishida: Language Grid: An Infrastructure for Intercultural Collaboration, SAINT-06, pp.96-100(2006).
[8] KODENSHA, http://www.kodensha.jp/
[9] NTT Natural Language Research Group, http://www.kecl.ntt.co.jp/icl/mtg/resources/index.php
[10] 会話表現データベース，ATR 音声翻訳通信研究所，http://www.atr-p.com/sdb.html