機械翻訳向け前編集に有効な書き換えルールに関する調査

宮田玲（東大/NICT），藤田篤（NICT）
rei@p.u-tokyo.ac.jp

1. 背景と目的

背景：

＊機械翻訳（MT）を実用的に活用するために前編集や制限言語が有効 (Pym, 1990; O'Brien&Roturier, 2007)

＊原文書き換えに関する様々な研究やガイドラインがあるが、その網羅性・有効性は十分明らかではない

どのような手順で、どのくらいの量・詳細度で書けば十分か？ (渡邊ほか, 2009; 小倉ほか, 2010; 松田, 2014)

目的：

＊書き換えルールの構築手順の検討

＊特定のドメイン（自治体、病院会話、新聞記事）、MTシステム（みんなの自動翻訳）、言語方向（日英）を対象とし、予備的な書き換えパターンの調査

2. 書き換えデータの収集

書き換え手法：

＊MTの品質向上を目指して人間が原文の書き換えを試行錯誤的に繰り返す (Miyata et al, 2015)

＋できるだけ最小単位の書き換え履歴を保存する

＋過去の書き換え履歴にいつでも戻れる

書き換えプラットフォーム：

進行状況
"In progress"
"Complete"
"Give up"

「ベスト」
全履歴中、MT品質が最も高い書き換え

「子」
当該要素から派生した書き換え

「親」
当該要素の派生元の書き換え

書き換え実験：

＊英語に熟練している日本語母語話者1名

＊100ユニット×3ドメイン=300ユニット

＊なるべく最小の書き換えごとに履歴を残す

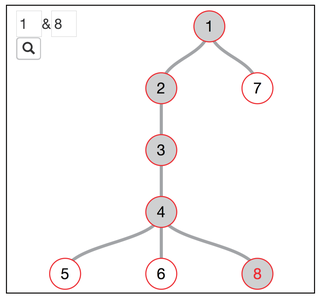
※1ユニット＝1つの原文（「オリジナル」）から派生する一連の書き換え履歴

	病院会話	自治体	新聞記事
原文の平均文長	20.2	34.8	44.4
総書き換え数	1199	2119	3823
Complete/Give up	97/3	97/3	86/14
オリジナル==ベスト	40	3	0

3. 書き換えデータの分析

分析対象：

（1）全書き換え履歴；（2）「オリジナル」から「ベスト」に至るパス；(3)「オリジナル」と「ベスト」のみ



出生届を居住地の市区町村の役所に提出	Birth registrations submitted to the public office of the municipality where you live.
出生届を居住地の市区町村の役所に提出してください	Please submit it to the city hall of the city, ward, town or village of the place of residence registration of a new birth
出生 届 を 居住 地 の 市区 町村 の 役所 に ご 提出 ください 。	Submit them to the office of city , ward , town or village of the place of residence birth certificates .
居住 地 の 市区 町村 の 役所 に 出生 届 を ご 提出 ください 。	Please submit birth registrations to the city hall of the city , ward , town or village of the place of residence .
お住まいの市区町村の役所に出生届をご提出ください。	Please submit notification of birth to the public office of the municipality where you live.

分析方法：「オリジナル」と「ベスト」を直接比較

出生 届 を 居住 地 の 市区 町村 の 役所 に 提出

お 住まい の 市区 町村 の 役所 に 出生 届 を ご 提出 ください 。

Birth registrations submitted to the public office of the municipality where you live .

Please submit notification of birth to the public office of the municipality where you live .

1. 表層表現の差異の抽出

「出生届を」の位置を変更
居住地→お住まい
提出→提出する
→ご提出ください

2. 書き換えパターンの同定

ヲ格を動詞の近くに移動
語彙の置換
体言止めに不足要素を追加
機能表現の追加

※分析対象ユニット
自治体ドメイン97ユニット
（「オリジナル==ベスト」の3ユニットは除外）

分析結果：

＊416の表層表現の差異

＊94の書き換えパターン

＊4つの大カテゴリー
- 構造
- 語彙（内容語、機能語）
- 表記
- その他（内容の変更など）

＊6の操作：追加、削除、置換、移動、分割、構造変更

書き換えパターンの例：

カテゴリー	操作対象	操作（詳細）	例
構造	読点	追加（節の区切り）	雨戸やシャッターがあれば、
構造	条件節	置換（特定表現の使用）	～すると→～する場合
構造	体言止め	追加（明示化）	開催。→開催されます。
語彙（内容語）	語彙	置換（具体化）	開ける→開放する
語彙（機能語）	機能表現	置換（特定表現の使用）	ましょう→ください
表記	ひらがな	置換（漢字化）	つとめる→努める
その他	内容	追加（具体化）	普通科→普通科コース

4. 結論と課題

＊効率的な書き換えパターン抽出のための手法とツールの提案

＊幅広い書き換えパターンの抽出（先行研究を大きく上回る）

1. 書き換えパターンの有効性の検証とルール化
・・・書き換えパターンの適用によるMTのパフォーマンス評価
2. 書き換えパターンの網羅性の検証
・・・ドメイン、データ量、MTシステム、言語方向、書き換え作業者

5. さらなる展望

＊書き換えの自動化
1. ルールベース書き換え
2. 統計ベース書き換え

＊分析手続きの明確化
・・・言語直観に依存しない

＊データ収集・分析コストの査定