2016年3月10日
言語処理学会第22回年次大会(NLP2016)
テーマセッション:文理・産学を超えた翻訳関連研究(1)

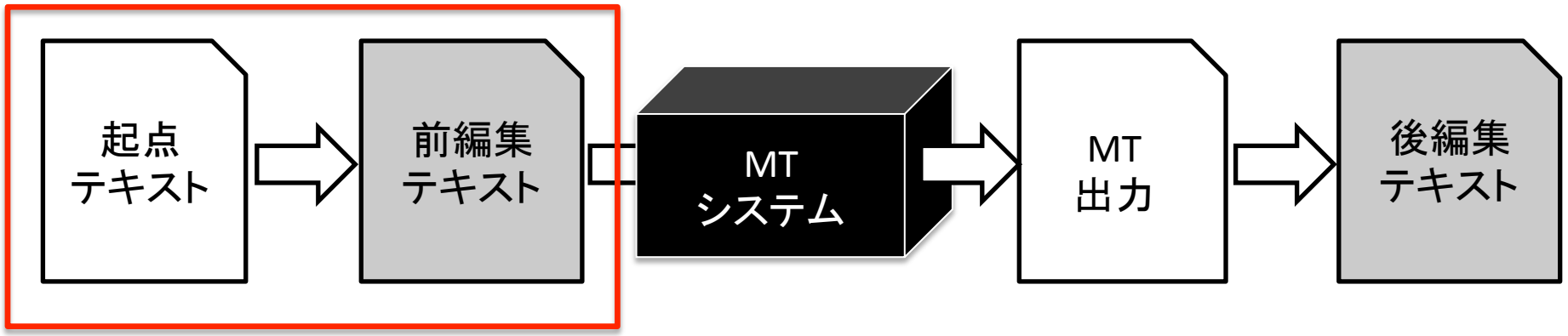
機械翻訳向け前編集の事例収集と類型化

宮田玲[†‡], 藤田篤[†], 内山将夫[†], 隅田英一郎[†]
[†]情報通信研究機構, [‡]東京大学

背景

- 機械翻訳(MT)性能の向上と社会的な応用
  - 商用・無料のMTの普及↔ブラックボックス
- 日英などの言語方向の翻訳は難しい

起点テキスト(原文)の統制によるMT品質の改善＝前編集



先行研究

- 目の前のテキストを翻訳しやすく書き換え
  - 人間による翻訳リペアとそのサポート
    (Uchimoto et al., 2006; Resnik et al., 2010; Miyabe et al., 2011)
  - 統計的前編集器の開発
    (Sun et al., 2010; 南條ほか, 2012)

書き換えの全体像を明らかにするものではない

先行研究

- 書き換えのパターンを列挙・分類
  - 制限言語・執筆ガイドライン
    (O'Brien et al., 2007; Hartley et al., 2012; Miyata et al., 2015)
  - 自動書き換えルール
    (Shirai et al., 1998; Mitamura & Nyberg, 2001)

- **異なるドメイン、言語方向、機械翻訳システムに対する網羅性・有効性は不明**
- **パターンを洗い出すための方法が確立されていない**

目的

- 機械翻訳向け前編集で見られる書き換え事例をドメイン横断的に収集し、類型化する

MTへの応用に向けた書き換えの現象理解

研究の流れ

1. 人手による前編集事例の収集
2. 収集した事例の分析・類型化

研究の流れ

1. 人手による前編集事例の収集
2. 収集した事例の分析・類型化

事例収集の手法

- 既存手法(Miyata et al., 2015)
  - 原文の内容を保持したまま、MT訳が十分な品質に達するまで、人間が書き換えを繰り返す
  - MT訳を見ながら、品質向上を目指して書き換える

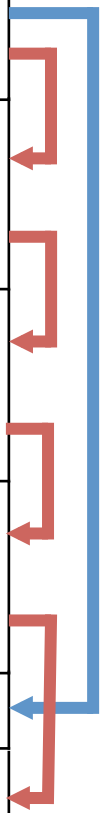
書き換えの試行錯誤の過程を細かく記録していない

提案手法の特徴

1. なるべく最小単位の書き換えを記録する
2. 過去の書き換え履歴に適宜戻ることを許す

書き換え手順

No	原文	MT結果
1	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.
2	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
3	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 certificate.
4	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.
	⋮	
8	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.



書き換えツール

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

十分な品質（＝多少流暢さに欠けても、情報の過不足がなく、文法的にも正しい訳文）に達したら、Complete
最後に、過去の書き換え履歴から、最もMT訳の品質が高いものを選ぶ（「ベスト」）

書き換え作業の実施

言語方向	日英
MTシステム	みんなの自動翻訳（汎用）
ドメイン（データセット）	病院内会話（**病院**） 自治体生活情報（**自治体**） 新聞記事 • 日本語原文（**BCCWJ**） • 英語原文日本語訳（**Reuters**）
データ	400ユニット （各データセット100ユニット）
作業者	日本語母語話者1名 （英語の翻訳評価が可能）

前編集事例の収集結果

データセット	平均文長 (標準偏差)	書き換え数				ユニット数	
		Total	Avg.	Med.	Max	原文=Best	Complete
病院	12.1 (4.5)	1199	12.0	3	105	40	97
自治体	21.3 (12.0)	2119	21.2	14.5	89	3	97
BCCWJ	26.9 (16.0)	3823	38.2	26.5	209	0	86
Reuters	34.8 (12.6)	5546	55.5	45	258	4	93

合計12,687の前編集事例を収集

病院ドメインの40/100ユニットは、最初から十分な翻訳品質を達成

90%以上のユニットで十分な翻訳品質を達成→MTのポテンシャル

十分な品質に達した例

ST 同国は、前年の過剰輸出と、今年の減産によって、穀物不足に直面しており、大量の小麦輸入の計画を表明している。

MT Excess exports in the previous year, and reduced production this year, is facing a shortage of grain, a large amount of wheat imports plan.

⬇ 194書き換え

ST 当年の減産と前年の過剰輸出による穀物の不足をふまえ、この国は小麦を大量に輸入する計画を表明している。

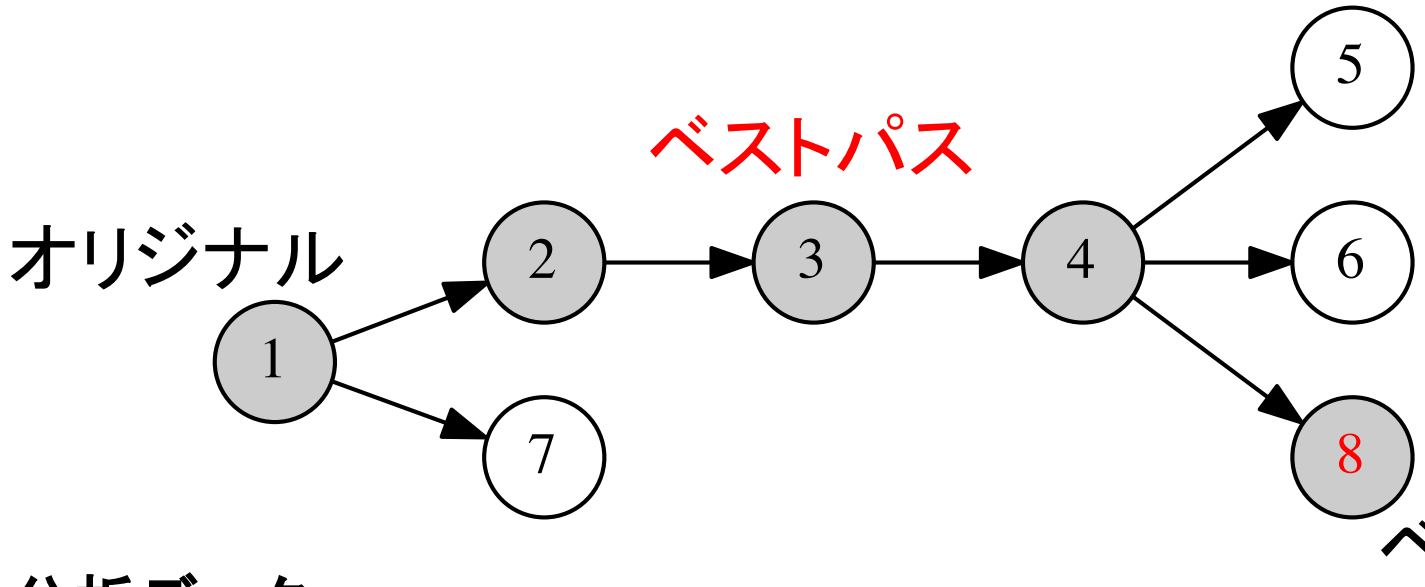
MT Based on the shortage of grain due to production cuts in the current year and excessive exports last year, this country has announced plans to import a large amount of wheat.

研究の流れ

1. 人手による前編集事例の収集
2. 収集した事例の分析・類型化

分析対象

オリジナルからベストに至る履歴中にMT訳の品質向上に寄与する前編集方法が必ず含まれている



分析データ:
ベストパス上の前編集事例
各データセット、10ユニット（合計40ユニット）

分析手続き

十日前後で登録のクレジットカードから引き落としを行います。

(1) 最小単位の書き換えへの分解

～で～から～→～から～で～	語順変更

(2) テキスト表層上の差異の抽出

登録のクレジットカードから十日前後で引き落としを行います。

(3) 書き換えの種類の類型化

～を行います→～が行われます	態の変更

登録のクレジットカードから十日前後で引き落としが行われます。

(1) 分解作業結果

各ドメイン10ユニットの書き換え事例数

データセット	全パス中の事例数	ベストパス中の事例数		
		(a)分解前	(b)分解後	(b)/(a)
病院	131	97	185	1.91
自治体	217	106	186	1.75
BCCWJ	484	174	340	1.95
Reuters	483	191	268	1.40

合計979書き換え事例
→類型化

(2)(3) 類型化結果

- 53種類
- 7カテゴリー
  - 構造：Structure
  - 語彙（内容語）: Lexicon (Content word)
  - 語彙（機能語）: Lexicon (Functional word)
  - 語彙（ターミノロジー）：Lexicon (Terminology)
  - 表記：Orthography
  - 内容：Information
  - エラー：Error

ID	書き換えの種類	事例の頻度			
		病院	自治体	BCCWJ	Reuters
S01	文分割／統合	4	1	7	2
S02	レイアウトの変更	0	3	0	0
S03	複文／重文の変更	0	0	0	1
S04	句の分割	0	0	1	0
S05	語順変更	24	6	22	13
S06	主語追加／削除	0	2	2	2
S07	読点の削除／追加	24	5	27	27
S08	主題のスコープ変更	0	0	1	1
S09	ガ格と主題ハの変更	0	1	3	2
S10	視点変更	0	2	11	0
S11	態の変更	3	1	13	3
S12	修飾の仕方の変更	2	0	12	13
S13	動詞句の主辞交替	0	0	0	3
S14	条件節の明示	2	7	2	0
S15	体言止めの回避／使用	0	1	3	5

ID	書き換えの種類	事例の頻度			
		病院	自治体	BCCWJ	Reuters
S16	名詞句の主辞交替	0	0	1	0
S17	名詞句・動詞句の交替	3	4	9	0
S18	複合動詞の使用／展開	2	0	2	0
S19	複合名詞の使用／展開	2	7	5	8
S20	接尾辞の使用／解除	2	1	10	5
S21	接続表現	6	16	12	13
S22	並列表現	2	3	1	0
S23	同格表現	0	0	0	5
S24	限定表現	0	0	0	3
S25	場所の限定表現	0	0	0	2
S26	伝聞表現	0	0	0	4
S27	間接疑問表現	0	0	0	1
S28	サ変名詞表現の変更	1	2	7	4
S29	形式名詞を使った表現	0	1	3	5
S30	存在動詞表現	1	0	0	1
S31	になる・となる表現	0	0	0	11

ID	書き換えの種類	事例の頻度			
		病院	自治体	BCCWJ	Reuters
C01	特定表現の使用	29	36	69	33
C02	具体化	5	3	2	1
C03	端的な表現の使用	0	5	0	0
C04	参照表現	0	0	0	1
C05	冗長化	0	1	0	1
F01	敬語化／非敬語化	19	11	14	4
F02	時制の変更	0	3	1	2
F03	並列語句のつなぎの表現	4	4	0	1
F04	助動詞の変更	1	0	0	0
F05	助詞の追加／削除／変更	4	9	24	9
F06	助詞の使用／回避	4	3	3	10
F07	複合助詞	0	1	1	5

ID	書き換えの種類	事例の頻度			
		病院	自治体	BCCWJ	Reuters
T01	固有表現	0	0	3	6
O01	表記の変更	1	7	7	4
O02	文末処理	0	1	2	0
O03	記号の追加／削除／置換	0	6	0	0
O04	省略の補完	0	0	3	2
O05	チャンキングの追加／削除	0	5	3	1
I01	内容の変更	18	20	27	16
I02	ニュアンスの変更	0	7	17	6
E01	表記ミス、文法ミス	3	1	4	6
E02	必要要素の削除／復元、原文にない情報の追加	19	0	6	6

頻出の書き換え種類

- C01「特定表現の使用」
  - 一度→一回
  - 習得する→学ぶ
- S05「語順変更」
- S07「読点の削除／追加」
- S21「接続表現」
- F01「敬語化／非敬語化」

ドメイン依存の書き換え種類

タイ農民銀行が売買代金で1位を獲得し、
2バーツ高の141バーツ。

新聞ドメイン（BCCWJ、Reuters）でよく見られる

- **S20「接尾辞の使用／解除」**
  - 2バーツ高の
  - 2バーツ上がり
- **S15「体言止めの回避／使用」**
  - 141バーツ。
  - 141バーツとなった。

多様な書き換え種類

- **S23「同格表現」**
  - ガッドゥム副総裁
  - 副総裁ガッドゥム
  - 副総裁であるガッドゥム
  - 副総裁のガッドゥム

新しい書き換え種類

- **S13「動詞句の主辞交替」**
  - 懸念を強め
  - 強い懸念を抱き
- **S10「視点変更」**
  - トイレットペーパーで山積みのカートを押す客
  - トイレットペーパーをカートに山積みにした客

cf.**言い換えのあれこれ**
http://paraphrasing.org/paraphrase.html

結論

- 人間による試行錯誤的な前編集の過程を、なるべく細かい単位で逐一記録する手法とツールを開発
  - 合計12,687の前編集事例を収集
  - 90%以上の原文は、MT訳が十分な品質を達成
- 事例の一部(979事例)を手作業で類型化
  - 7カテゴリー・53種類の書き換えを同定
  - 特定ドメインに頻出の書き換え種類を把握
  - これまでに報告されていない種類を発見

今後の課題

- 作業者へのインストラクションとツールの改良
  - 「最小単位」の書き換えの促進
  - 直線的な「行ったり来たり」の予防
- 異なる言語方向、作業者、MTシステムでの比較
  - 現在、英日翻訳タスクでのデータ収集が完了
  - 作業者も日英・英日ともに2名に増やして、検証中

展望

- 前編集用書き換えルールの作成と自動適用
  - 類型化した書き換え種類の中から、特にMT品質の向上に寄与するものを取り出す (Miyata et al., 2015)
  - 自動的に適用できるかの見極め
- 統計的前編集器の開発 (Sun et al., 2010; 南條ほか, 2012)
  - 収集した前編集事例はモデルの訓練に使える
  - データをどのように使うかが重要

ありがとうございました