

人間の翻訳リペア

2009年

過程の分析

9月 30日

渡邊 裕之, 藤田 篤, 松原 仁

 公立はこだて未来大学

概要

翻訳リペアとは、正しい翻訳結果が得られるように
入力文を書き換える行為

課題：翻訳リペアには多大なコストを伴う

目標：翻訳リペアの自動化による翻訳支援

- 人間の折り返し翻訳を用いた翻訳リペア方法の調査
 - 翻訳リペアタイプの分布

背景と目的

背景

- 異文化コミュニケーション
 - 言語の壁
 - 多言語を十分に習得するのは容易ではない
- 言語グリッド[NICT, '06]
 - インターネット上の言語資源(機械翻訳, 対訳辞書等)のサービス基盤
- 機械翻訳
 - 正確な翻訳は難しい
 - 翻訳リペアによって訳質を上げることができる[小倉ら, '05]

目的

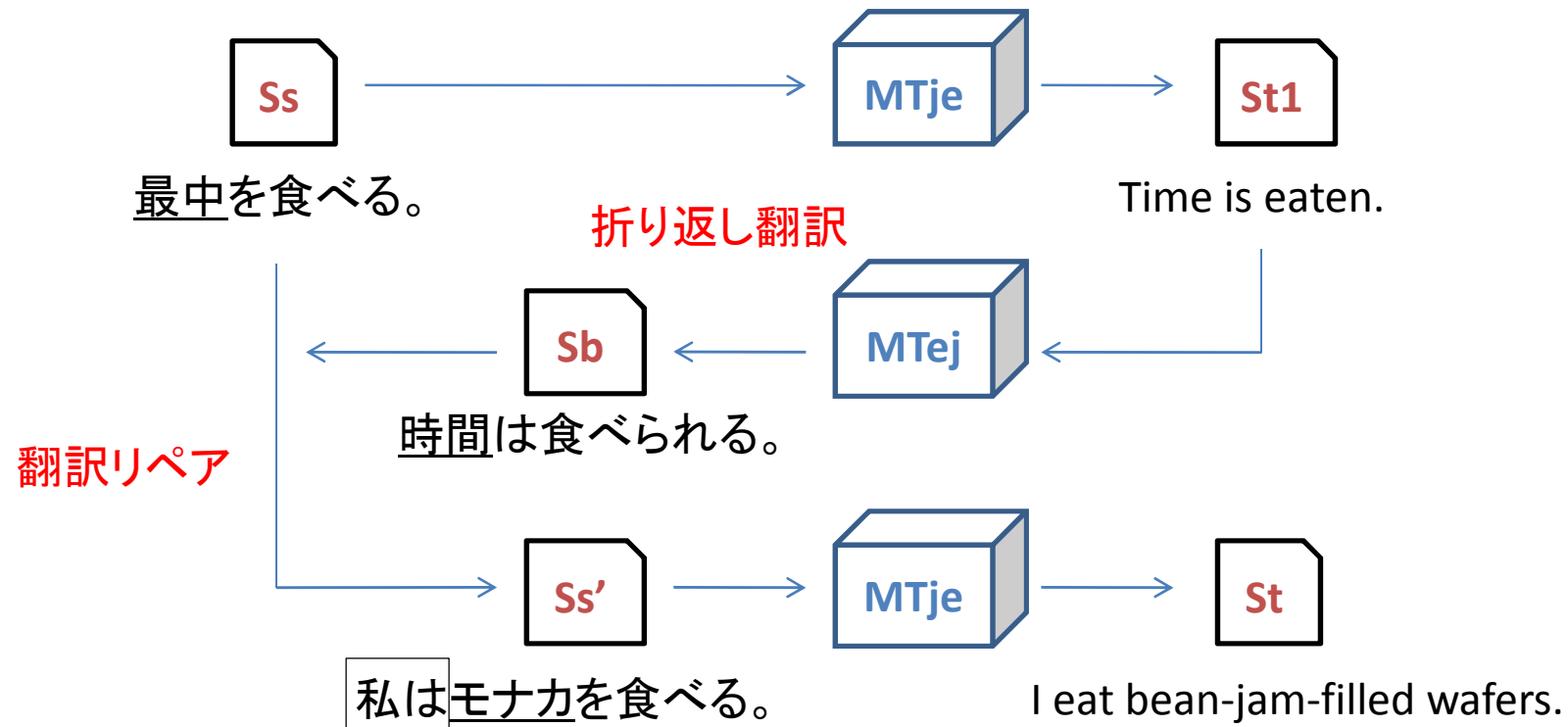
1. 人間の翻訳リペア過程の分析
2. 翻訳リペアの自動化

[NICT, '06] NICT言語グリッドプロジェクト, <http://langrid.nict.go.jp/>

[小倉ら, 05]小倉健太郎, 林田良彦, 野村早恵子, 石田享, “機械翻訳を介したコミュニケーションにおける利用者の機械翻訳システム適応の言語依存性”, 自然言語処理, Vol.12, pp.183-201, 2005.

折り返し翻訳を用いた翻訳リペア [山下ら, '06]

翻訳リペア: 正しい翻訳結果が得られるように入力文を書き換える行為



- 母国語の知識しか持たないユーザでもリペア可能

[山下ら, '06] 山下直美, 坂本知子, 野村早恵子, 石田享, 林良彦, 小倉健太郎, 井佐原均, “機械翻訳へのユーザの適応と書き換えへの教示に関する分析”, 情報処理学会論文誌, Vol.47, No. 4, 2006

課題

- リペアには多大な作業コストを伴う
 - リペア回数, 作業時間
 - 不慣れな程, リペアに苦戦する



リペアの自動化による翻訳支援

アプローチ

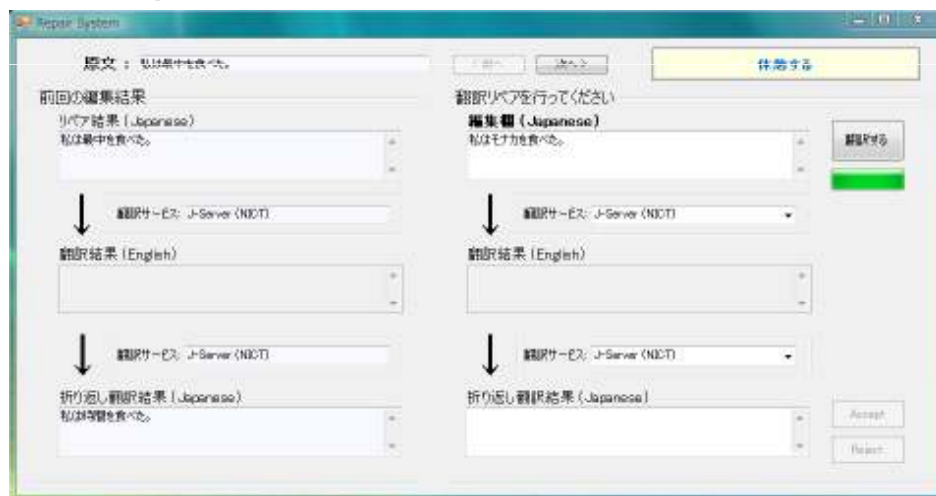
1. (人間の)翻訳リペア方法の調査

- リペアしていく過程のデータを収集
- 自動化が有効なリペアの種類を特定
- リペアの優先度から自動リペアの戦略を検討

2. 翻訳リペアの自動化

データ収集の環境

- 被験者
 - 大学院生・大学生合計5人
 - 翻訳リペアについての説明を受けた
- 作業用ツール



機械翻訳サービスは高電社のJ-Serverを言語グリッドを介して使用

- 入力文
15～40文字の日本語文100文
 - 機械翻訳機能試験文
[池原ら, '94]から抜粋

3,718文
↓ 15～40文字
1,834文
↓ サンプル
100文

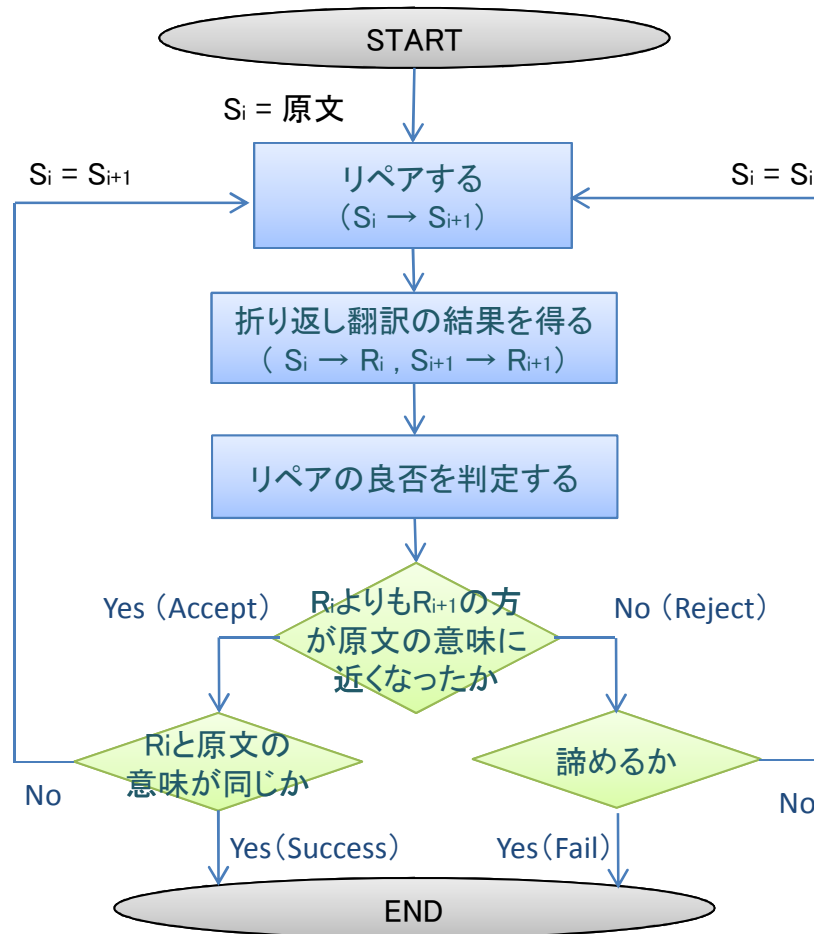
- 長さの制約
簡単な内容の文ではなく、ある程度表現可能な文字数
[宮部ら, '07]

[池原ら, '94]池原悟, 白井諭, 小倉健太郎, “言語表現体系の違いに着目した日英機械翻訳機能試験項目の構成”, 人工知能学会誌, Vol.9, No.4, pp.569-579, 1994










[宮部ら, '07]宮部真衣, 吉野孝, 重信智宏, “折り返し翻訳を用いた翻訳リペアの効果”, 電子情報通信学会論文誌, Vol.J90-D, No.12, pp.3141-3150, 2007

リペア事例の収集

翻訳リペアの作業手順



収集するデータ

-  被験者 (例: 被験者A)
-  リペア前の文 (例: モナカを食べる。)
-  翻訳結果 (例: I eat bean-jam-filled wafers.)
-  機械翻訳の種類 (高電社のJ-Server)
-  リペア時間 (例: 30秒)
-  原文 (例: 最中を食べる。)
-  リペア後の文 (例: 私はモナカを食べる。)
-  折り返し翻訳の結果 (例: 時間は食べられる。)
-  折り返し翻訳で用いた機械翻訳の種類 (高電社のJ-Server)
- Accept / Reject
折り返し翻訳の結果が原文の意味に近づいたか

分析

リペアなしに Success とした文数

被験者	A	B	C	D	E	平均
文数	14	17	14	16	19	16

Success/Fail の分布

被験者	A	B	C	D	E	総計
Success	93	89	86	95	91	454
Fail	7	11	14	5	9	46

平均ターン数

被験者	A	B	C	D	E	総計
Success	8.7	5.5	6.9	4.4	4.2	5.9
Fail	5.1	8.5	14.9	7.4	5.2	9.2
総計	8.4	5.8	8.0	4.5	4.3	6.2

ターン数

被験者	A	B	C	D	E	総計
accept	281	127	230	257	118	1013
reject	463	356	474	97	215	1605
総計	744	483	704	354	333	2618

文字数の変化(平均)

被験者	A	B	C	D	E	総計
accept	-0.40	0.31	0.35	0.04	-1.01	-0.14
reject	-0.04	0.49	0.98	0.63	-0.44	0.32

結果

1. 単語単位の書き換え

・動詞の言い換え

リペア前の文

それは真実であることが**わかった**。

リペア後の文

それは真実であることが**判明した**。

2. 文中の要素を追加

・目的語の明記

リペア前の文

私が駅に着くと、彼は迎えに来ていた。

リペア後の文

私が駅に着くと、彼は**私を**迎えに来ていた。

3. 文中の要素を削除

・不要な語句の削除

リペア前の文

この会談で両者の食い違いがさらに広がった**形となった**。

リペア後の文

この会談で両者の食い違いがさらに広がった。

リペアの自動化に向けて

- 以下のルールに基づくリペアの可否を検討中
 - 特定の表現を書き換える
 - 「ことがある」→「可能性がある」
 - 「こともある」→「場合もある」
 - ひらがなを漢字にする
 - 「川ざかな」→「川魚」
 - 「とめる」→「止める」
 - 特定の表現を含む文へ副詞を追加する
 - 「～したことがない」を含む文→「一度も」を追加
 - 「～だけが」を含む文→「唯一の」を追加
 - 特定の表現を削除する
 - 「というもの」、「こととする」、「形となった」

翻訳結果の精度を確認する必要がある
今回とは別の入力文でも試す必要がある

Acceptされた翻訳リペアタイプの分布(重複あり)

被験者 リペアタイプ	A	B	C	D	E	総計
1. 単語単位の書き換え	116	86	130	138	66	536
2. 文の要素を追加	58	14	48	39	7	166
3. 文の要素を削除	48	9	15	19	24	115
4. 語順変更	22	5	10	25	8	70
5. 文の分割	15	2	0	8	3	28
6. 態の交替	4	0	3	2	1	10
7. 文の統合	5	0	0	2	0	7
8. 固有表現のチャンキング	0	0	1	2	2	5
9. その他	47	19	29	35	11	141
総計	315	135	236	270	122	1078

4. 語順変更

・主語の移動

リペア前の文	その理由は鯨がエラを持たないことです。
リペア後の文	鯨がエラを持たないことがその理由です。

5. 文の分割

リペア前の文	これまでに応募者は80人と既に定員の倍以上になっています。
リペア後の文	これまでに応募者は80人です。既に定員の倍以上になっています。

6. 態の交替

・受動態→能動態

リペア前の文	テーブルが新しいのに置き換えられた。
リペア後の文	テーブルを新しいのに置き換えた。

8. 固有表現のチャンキング

リペア前の文	この列車はブルートレインと言う。
リペア後の文	この列車は”ブルートレイン”と言う。

今後の計画

- 折り返し翻訳を用いた翻訳リペア方法の分析
 - Rejectされた翻訳リペアタイプの分布
- 翻訳結果を用いた翻訳リペアの調査
- 翻訳リペア自動化の検討
 - ルールに基づく翻訳リペアの検討