# Dissecting Human Pre-Editing toward Better Use of Off-the-Shelf Machine Translation Systems

Rei Miyata (Nagoya University; miyata@nuee.nagoya-u.ac.jp)  Atsushi Fujita (NICT)

## 1. Objective

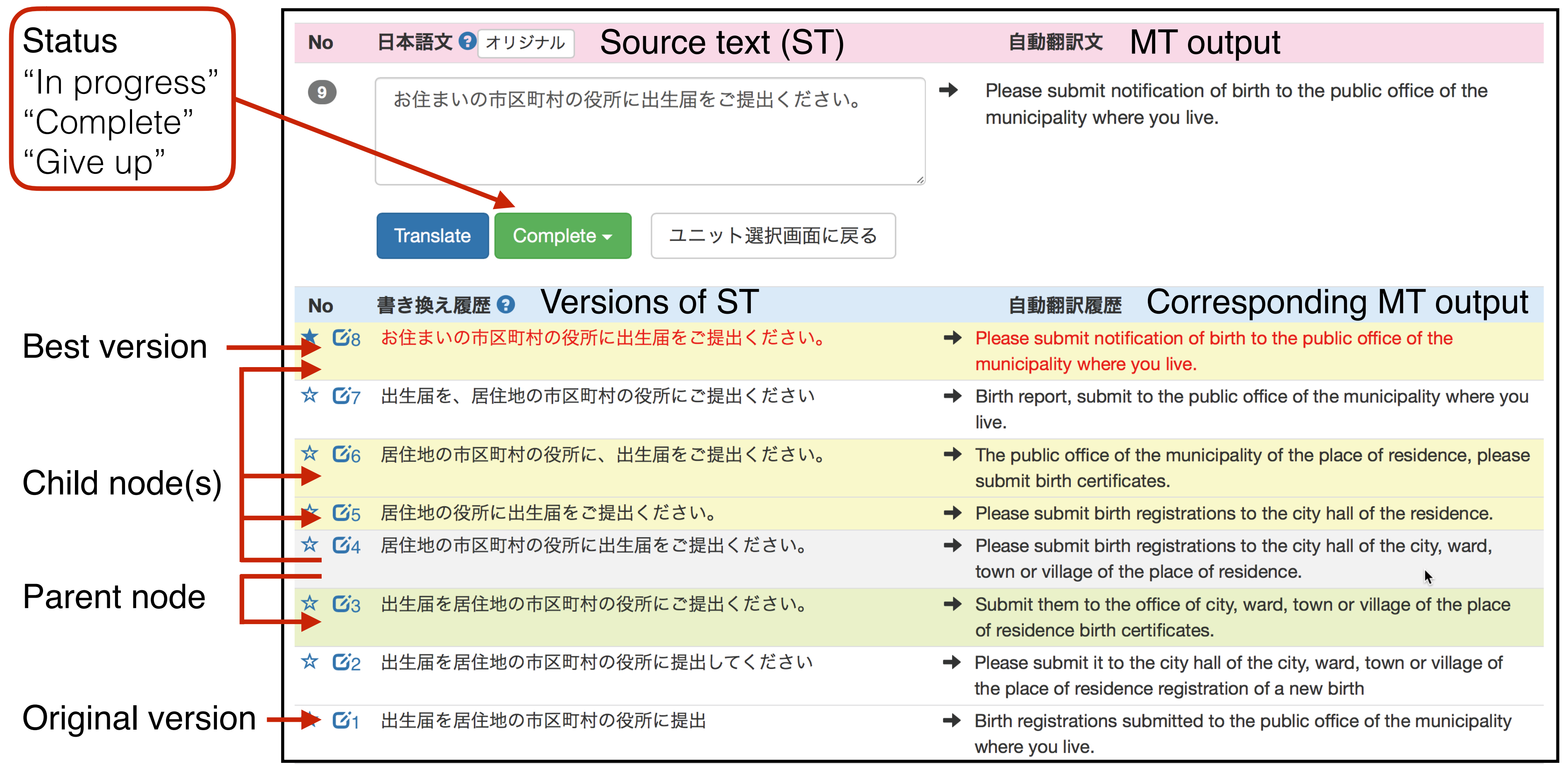
1) Investigate the **capability of the pre-editing strategy**
- Design a human-in-the-loop protocol to collect pre-edit instances
- Collect pre-edit instances in Japanese-to-English translation tasks on 4 datasets

2) Provide an **overview of possible edit operations**
- Create a typology of edit operations

## 2. Protocol for Collecting Pre-Edit Instances

Human editors incrementally edit source texts (STs) relying on their introspection, so that improved MT quality is achieved. **(Miyata et al. 2015)**

**New features:**
- Record ST after every minimal edit is performed
- Allow editors to resume editing from any given past versions of ST



Status
"In progress"
"Complete"
"Give up"

Best version

Child node(s)

Parent node

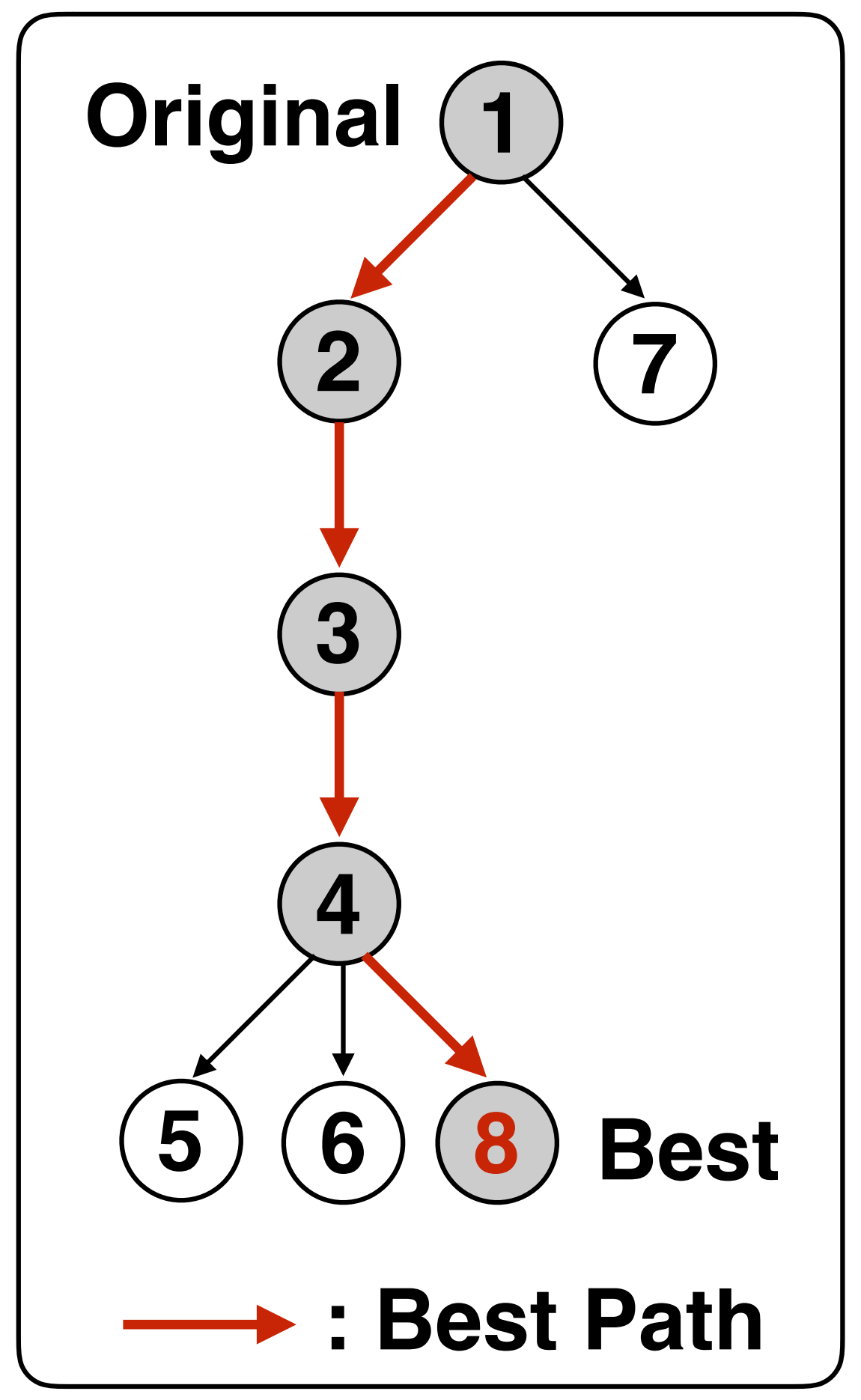
Original version

| No | 日本語文 オリジナル Source text (ST) | | 自動翻訳文 MT output |
|---|---|---|---|
| 9 | 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. |

Translate | Complete ▾ | ユニット選択画面に戻る

| No | 書き換え履歴 Versions of ST | | 自動翻訳履歴 Corresponding MT output |
|---|---|---|---|
| ★ 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. |
| ☆ 7 | 出生届を、居住地の市区町村の役所にご提出ください | ➜ | Birth report, submit to the public office of the municipality where you live. |
| ☆ 6 | 居住地の市区町村の役所に、出生届をご提出ください。 | ➜ | The public office of the municipality of the place of residence, please submit birth certificates. |
| ☆ 5 | 居住地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 residence. |
| ☆ 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. |
| ☆ 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 certificates. |
| ☆ 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 |
| ☆ 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. |

***Tree representation of versions of STs in a 'unit'***



Original

Best

➜ : Best Path

| | |
|---|---|
| **Step 1** | Assess the MT output for the present ST on a 5-point scale criterion (5: Perfect — 1: Incorrect/nonsense). Go to **Step 4**, if it has satisfactory quality (>=4: Good); otherwise, go to **Step 2**. |
| **Step 2** | Select one version of ST to be edited, and go to **Step 3**; if none is likely to achieve satisfactory quality, go to **Step 4**. |
| **Step 3** | Minimally edit the selected version of ST, while keeping the meaning of the ST. Go back to **Step 1**. |
| **Step 4** | Choose one version of ST that achieves the best MT quality (Best ST). |

## 3. Pilot Run

- 4 datasets: hospital conversation (**hosp**), municipal information (**muni**), 2 news articles (**bccwj**: Japanese-origin & **reuters**: English-origin)
- MT system: **TexTra** (freely-available, state-of-the-art phrase-based SMT)
- A Japanese editor with ample experience in evaluating MT quality

***Statistics of the collected data***

**More than 85% of the STs achieved satisfactory quality**

| Dataset | Mode | Avg. num. of tokens in original ST (s.d.) | Num. of versions of ST | | | | Num. of units | |
|---|---|---|---|---|---|---|---|---|
| | | | Total | Avg. | Med. | Max | Original=Best | Satisfactory quality |
| *hosp* | spoken | 12.1 (4.5) | 1199 | 12.0 | 3 | 105 | 40 | **97** |
| *muni* | written | 21.3 (12.0) | 2119 | 21.2 | 14 | 89 | 3 | **97** |
| *bccwj* | written | 26.9 (16.0) | 3823 | 38.2 | 26 | 209 | 0 | **86** |
| *reuters* | written | 34.8 (12.6) | 5546 | 55.5 | 45 | 258 | 4 | **93** |

***Example of Best ST with satisfactory MT quality***

| | ST | MT output |
|---|---|---|
| **Original** | 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。 | Excess exports in the previous year, and reduced production this year, is facing a shortage of grain, a large amount of wheat imports plan. |
| **Best** | 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。 | 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. |
| **Reference** | The country, currently battling an acute grain shortage due to excessive exports last year, faces a poor harvest this year and intends to import large quantities of wheat. | |

**25 edits**

## 4. MTranslatability for Different Languages

- ST: **Original** vs. **Best**
- TL: **Chinese** & **Korean**
- MT: **TexTra**
- Criterion**: 5-point scale**

***Human evaluation of MT quality (Chinese & Korean)***

| Chinese | Avg. score | | Num. of units (Org vs. Best) | | |
|---|---|---|---|---|---|
| | Org | Best | > | = | < |
| *hosp* | 2.73 | **2.93** ** | 7 | 70 | **23** |
| *muni* | 2.84 | 2.89 | 32 | 31 | 37 |
| *bccwj* | 2.39 | **2.75** ** | 13 | 42 | **45** |
| *reuters* | 2.61 | 2.77 | 22 | 45 | 33 |

| Korean | Avg. score | | Num. of units (Org vs. Best) | | |
|---|---|---|---|---|---|
| | Org | Best | > | = | < |
| *hosp* | 3.32 | **3.56** ** | 12 | 57 | **31** |
| *muni* | 3.58 | 3.67 | 32 | 29 | 39 |
| *bccwj* | 3.37 | **3.60** * | 18 | 47 | **35** |
| *reuters* | 3.31 | 3.36 | 24 | 47 | 29 |

## 5. Typology of Edit Operations

- **Best path** of 10 randomly sampled units

= 979 instances of **primitive edit operations**

⬇

- 53 types of **edit operations** in 7 major categories:
  **S** (**S**tructure), **C** (**C**ontent word), **F** (**F**unctional word), **T** (**T**erminology), **O** (**O**rthography), **I** (**I**nformation), **E** (**E**dit that causes/resolves error in ST)

***Less frequent type (not covered by related work)***

**S13: Head-switching of verb phrase**

| | |
|---|---|
| 懸念を強め | (strengthen anxiety) |
| → 強い懸念を抱き | (have strong anxiety) |

***Frequent types***

**C01: Alternative lexical choice**

一度 → 一回 (once)

習得する → 学ぶ (learn)

**S05: Phrase reordering**

**S07: Insertion/deletion of punctuation**

***Domain specific types (e.g. news domain)***

**S15: Use/disuse of clause-ending noun**

**S20: Use/disuse of nominal/verbal suffix**

前月比で → 前月に比べて
(compared to the previous month)

## 6. Conclusion and Future Work

1) Collected 12,287 pre-edit instances
- High capability of the off-the-shelf MT system
- Improved Chinese and Korean translatability

2) Built a typology of a wide range of edit operations

**Develop an automatic pre-editor**
- Controlled language formulation by assessing the effectiveness of each type of edit operation
- Statistical model based on our collection of pre-edit instances