# Understanding Pre-Editing for Black-Box Neural Machine Translation

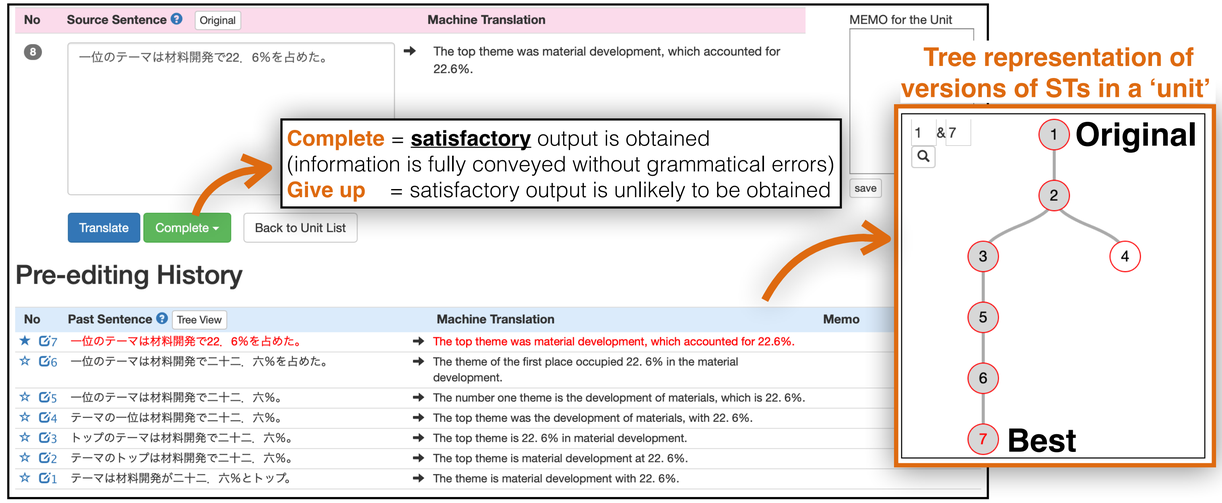
Rei Miyata (Nagoya University; miyata@nuee.nagoya-u.ac.jp)  Atsushi Fujita (NICT)

## 1. Objectives

1. Investigate the **capability of the pre-editing + black-box NMT** through a human-in-the-loop protocol for collecting pre-editing instances
2. Analyse the instances to understand the **diversity and impact of pre-editing**

## 2. Protocol for Collecting Pre-Edit Instances

Incrementally and minimally rewrite an ST on a trial-and-error basis with the aim of obtaining better MT output [Miyata & Fujita 2017]




## 3. Collected Instances

- **3 translation directions**: Ja-En, Ja-Zh, Ja-Ko
- **2 MT system**: Google, TexTra
- **100 STs in Japanese** from **4 domains**
  (hospital conversation, municipal procedure, Japanese-origin news article, English-origin news article)
- One **professional translator** for each translation direction

| Lang. | System | # instances Total | # instances Avg. | # satisfactory versions Orginal | # satisfactory versions Best |
|---|---|---|---|---|---|
| Ja-En | Google | 1332 | 13.32 | 27/100 | 98/100 |
| | Textra | 1260 | 12.60 | 23/100 | 96/100 |
| Ja-Zh | Google | 1371 | 13.71 | 2/100 | 91/100 |
| | Textra | 812 | 8.12 | 8/100 | 94/100 |
| Ja-Ko | Google | 950 | 9.50 | 1/100 | 96/100 |
| | Textra | 927 | 9.27 | 8/100 | 96/100 |
| Total | | 6652 | 11.09 | 69/600 | 571/600 |

**Attainable accuracy of pre-editing + NMT**

## 4. Analyses

### *a. Characteristics of pre-edited sentences*

- **Structural complexity ↑**
- **Low-frequency words ↓**

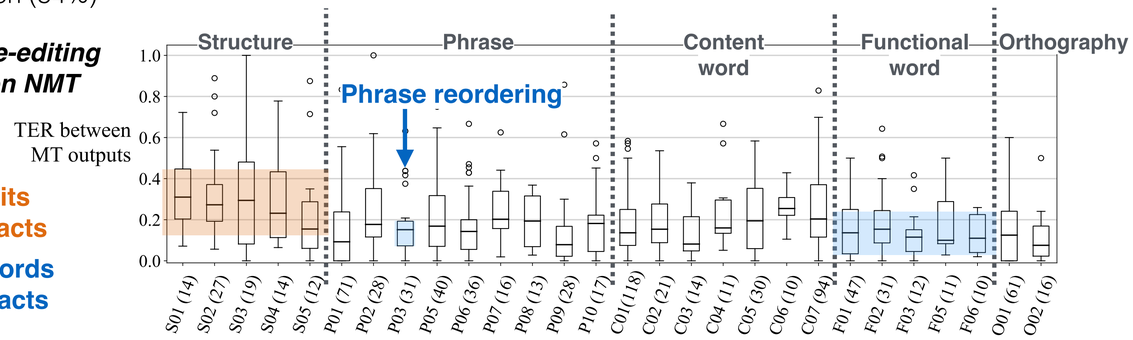
| | Org | Best Ja-En Google | Best Ja-En Textra | Best Ja-Zh Google | Best Ja-Zh Textra | Best Ja-Ko Google | Best Ja-Ko Textra |
|---|---|---|---|---|---|---|---|
| Sentence length | 25.4 | 27.8 | 26.9 | 28.6 | 27.1 | 27.8 | 26.9 |
| Attachment distance | 1.95 | 1.97 | 1.99 | 1.99 | 1.99 | 2.00 | 1.98 |
| Dependency depth | 3.57 | 3.73 | 3.68 | 3.73 | 3.77 | 3.78 | 3.76 |
| Type/Token Ratio | 0.398 | 0.386 | 0.395 | 0.387 | 0.392 | 0.384 | 0.392 |
| Word freq. rank (Med.) | 170 | 143 | 154 | 143 | 155 | 143 | 169.5 |

### *b. Diversity of pre-editing operations*

- **39 pre-editing operation types** under **6 categories**
  - Structure / Phrase / Content word / Functional word / Orthography / Error
- Classification based on the **information strategies**
  - **Explicitation (37%)**
  - Implicitation (10%)
  - Preservation (54%)

| Subcategory | # | Example of ST pre-editing | MT output |
|---|---|---|---|
| Information addition | 142 | 12日は台湾の休日のため休場。 | The twelfth is a holiday in Taiwan. |
| | | → 12日は台湾の休日のため**株式市場は**休場。 | **The stock market** was closed on the twelfth due to a holiday in Taiwan. |
| Use of clear relation | 103 | 来院しなくても10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。** | **Withdraw from your registered credit card** in about 10 days without visiting the hospital. |
| | | → 来院しなくても10日前後で登録のクレジットカードから**引き落としが行われます。** | Even if you do not visit the hospital, **your credit card will be debited** in about 10 days. |
| Use of narrower sense | 54 | 採尿と採便を**出して**ください。 | Please **collect** urine and feces. |
| | | → 採尿と採便を**提出して**ください。 | Please **submit** urine and stool samples. |
| Normalisation | 30 | 単位は億円。 | Figures are in **billions of yen**. |
| | | → 単位は億円**です**。 | The unit is **100 million yen**. |

### *c. Impact of pre-editing operations on NMT*




- **Structural edits → Major impacts**
- **Functional words → Minor impacts**

## 5. Findings

- **Enhancing the explicitness of an ST is more important** than making the ST shorter and simpler
- Although NMT is generally unpredictable, **there are recognisable tendencies of impact of pre-editing on NMT**

## 6. Future Application

- Tools to support human pre-editors
- Automatic pre-editors for black-box NMT
- Robust NMT models using our collected data