LREC2022

# A Benchmark Dataset for Multi-Level Complexity-Controllable Machine Translation

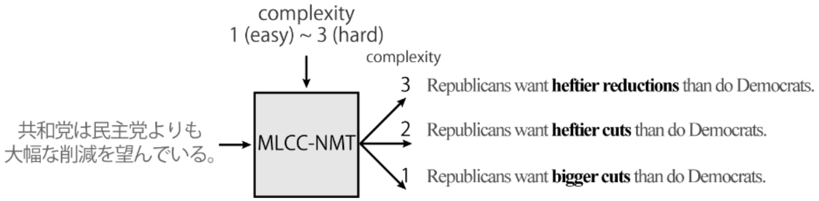
Kazuki Tani[1], Ryoya Yuasa[1], Kazuki Takikawa[2], Akihiro Tamura[1], Tomoyuki Kajiwara[2], Takashi Ninomiya[2], Tsuneo Kato[1] ([1]Doshisha University, [2]Ehime University)

## Introduction

□ Style-controllable NMT has recently received much attention.

□ Multi-Level Complexity-Controllable MT (MLCC-MT)

- ➢Controls the complexity of a target language sentence at three or more levels
- ➢To allow translation tailored to the user's reading level


complexity
1 (easy) ~ 3 (hard)
complexity
共和党は民主党よりも
大幅な削減を望んでいる。
MLCC-NMT
3 Republicans want **heftier reductions** than do Democrats.
2 Republicans want **heftier cuts** than do Democrats.
1 Republicans want **bigger cuts** than do Democrats.

**Problem:** Existing test dataset cannot precisely measure model performance

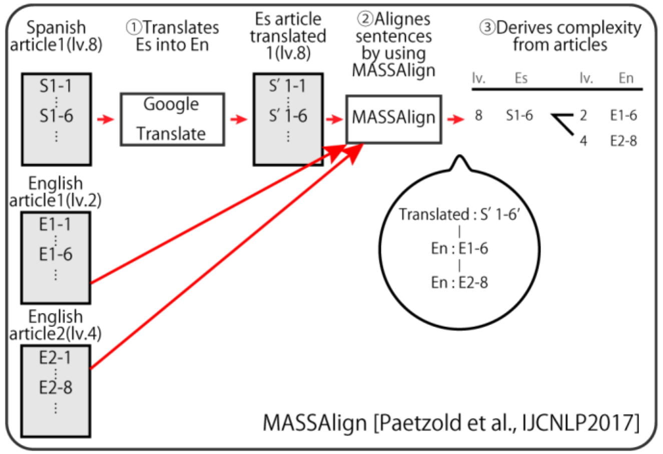
□ **Objectives:**

- ➢**Construct a test dataset to properly evaluate MLCC-MT models**
- ➢**Provide benchmark performance by evaluating two MLCC-MT models (i.e., pipeline and multi-task models)**

## Problems of Existing Test Dataset [Agrawal et al., EMNLP 2019]

□ Es⇔En dataset automatically generated from Newsela corpus

- ➢ Newsela corpus:
  - a document-level comparable corpus with document-level complexity (i.e., grade level), composed of English and Spanish news articles.
- ➢ Creation procedure:



MASSAlign [Paetzold et al., IJCNLP2017]

□ **Issues:**

1. **Incorrect translation pairs**
2. **Difference of granularity of information among target language sentences with different complexity levels**

Examples of Issue 2 (proper noun insertion)

| level | English |
|---|---|
| 8 | However, she says that there are times when "you just need to get away." |
| 5 | Yet she says that there are times when "you just need to get away." |
| 3 | Still, **Bopp** says that there are times when "you just need to get away." |

3. **Incorrect sentence-level complexity**

Examples of Issue 3

| level | English |
|---|---|
| 12 | So few Indians drink brewed coffee that virtually all its best crop is exported to countries such as Italy, where the beans are used in name-brand espresso blends and sold at a huge markup. |
| 9 | **There the beans are used in name-brand espresso blends and sold at a huge price increase.** |
| 7 | **There, the beans are used in name-brand espresso blends and sold for a huge price increase.** |

## Benchmark Test Dataset

□ A new benchmark test dataset for Ja-to-En MLCC-MT

□ Proposed creation procedure:

Step1: Generates the sets of English sentences with the same content written at multiple complexity levels

1-1: Extracts English aligned sentences from Newsela-auto [Jiang et al, ACL 2020]

1-2: Automatically removes exactly the same sentences or grade levels diff ≦ 1 (→ solves Issue 3)

1-3: Manually removes the sets where new content appears (→solves Issue 2)

Step2: Manually translates English sentences into Japanese (→solves Issue 1)

□ Result: **1,014 sets** (1 Japanese ⇔ 3～5 English with multi-level complexity)

□ Release: https://github.com/K-T4N1/A-BenchmarkDataset-for-ComplexityControllableNMT

Samples in our test dataset

| Japanese | level | English |
|---|---|---|
| 公衆衛生について の入門書が必要だ。 | 12 | A primer about public health is in order. |
| | 9 | A short explanation about public health is in order. |
| | 6 | A short explanation about public health is needed. |

## Benchmark Experiments

□ Implement two Transformer-based MLCC-MT models and evaluate them on our test dataset to serve as benchmark performance for future research

□ Benchmark models:

- ➢ Pipeline model: Ja-to-En NMT → En multi-level simplification
  - Ja-to-En NMT: Transformer NMT [Kiyono et al., wmt2020]
  - En multi-level simplification model: Incorporation of special tokens representing target complexity [Scarton et al., ACL 2018]
- ➢ Multi-task model: based on the following three losses

$$loss = L_{MT} + L_{Simplify} + L_{CMT}$$

- $L_{MT}$: the loss for conventional MT
- $L_{Simplify}$: the loss for text simplification
- $L_{CMT}$: the loss for complexity-controllable MT

□ Evaluation Metrics:

- ➢BLEU [Papineni et al., ACL 2002]: MT performance
- ➢SARI [Xu et al., TACL 2016]: text simplification performance
- ➢MAE$_{fkgl}$ (Mean absolute error of FKGL) [Nishihara et al., ACLSRW 2019]: complexity controlling performance

□ Results:

| model | BLEU (%) ↑ | SARI (%) ↑ | MAE$_{fkgl}$ ↓ |
|---|---|---|---|
| Pipeline | 15.12 | 23.89 | 5.10 |
| Multi-task | **20.17** | **26.78** | **4.83** |

➢Multi-task model outperforms the pipeline model in term of BLEU, SARI and MAE$_{fkgl}$.

## Conclusion

□ Create a new benchmark **test dataset** for Ja-En MLCC-MT

➢The proposed creation procedure includes **automatic filtering**, **manual check**, and **manual translation** to make our test dataset more appropriate than existing test datasets .

□ Implement two Transformer-based MLCC-NMT (pipeline and multi-task) and evaluate them as **benchmark performance**

□ Future work:

➢Increase the size of our dataset and create a multi-lingual dataset for MLCC-MT