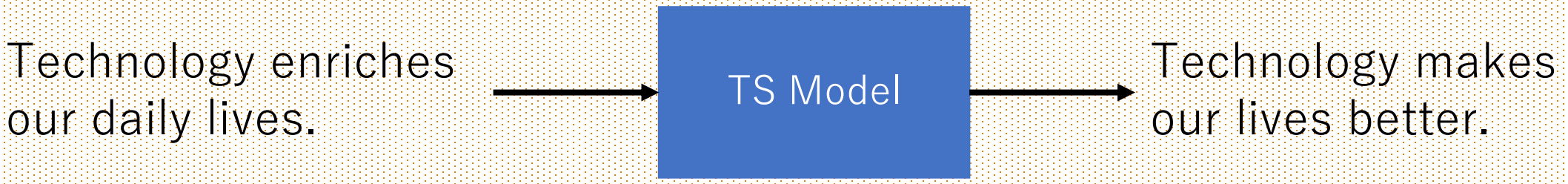
# Transfer Fine-tuning for Quality Estimation of Text Simplification

Yuki Hironaka[1], Tomoyuki Kajiwara[1], Takashi Ninomiya[1]

[1]Graduate School of Science and Engineering, Ehime University, Japan

# Background : Text Simplification (TS)

- Task that paraphrases complex expressions into simpler ones while preserving their meaning.




- Effect of TS
  - Contribute to learning and reading support for children and language learners.
  - Improve the performance of other NLP tasks.
    Ex.) Relation Extraction, Machine Translation

# Background : Evaluation of TS

## Automatic

Metrics
(BLEU, SARI)

(Input)
System Output
Reference

Evaluate word agreement between system output and reference

## Human

Input
System Output
(Reference)

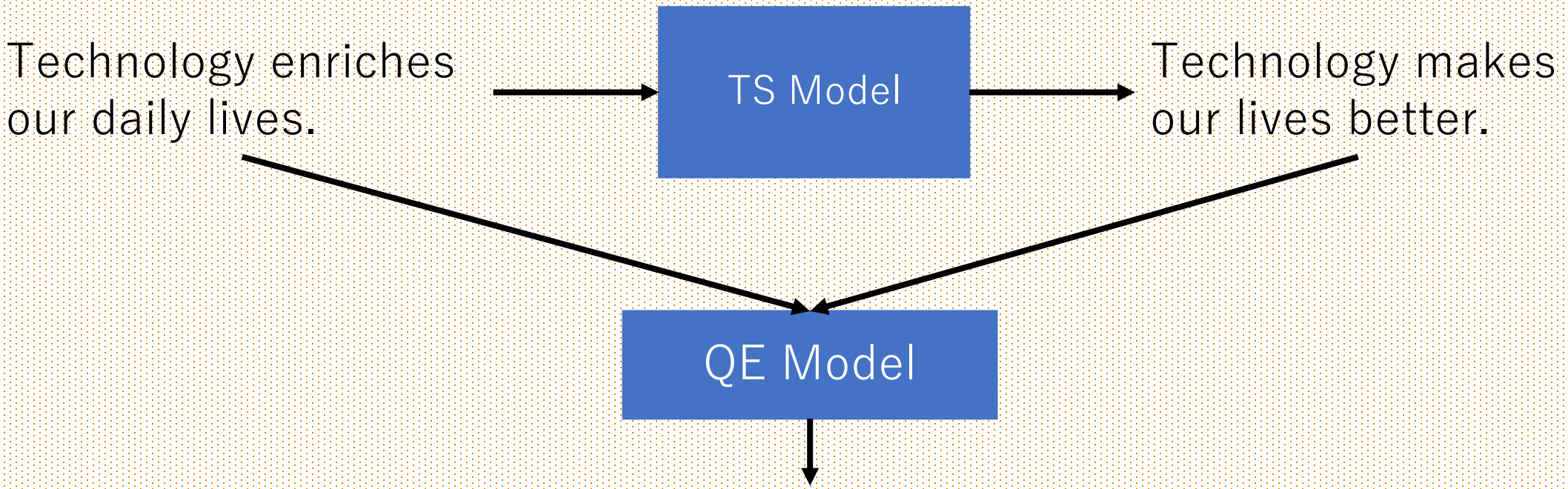
Evaluate in terms of fluency, meaning preservation, and simplicity

| | Cost | Reliability |
|---|---|---|
| Automatic | Low | Low |
| Human | High | High |

**Quality Estimation for TS** is being studied for low-cost and high-reliability evaluations.

# Background : Quality Estimation (QE) for Text Simplification

Task of estimating the quality of the output sentences from the input and output sentence pairs.



# Related Work : QE for TS

- Dataset
  - QATS [Štajner+ 2016]
    - Consisted 631 English sentence pairs.
    - Targeted models based on statistical machine translation.
    - Manually evaluated on a scale of Good, OK, and Bad for 4 aspects: fluency, meaning preservation, simplicity, and overall evaluation.
  - Simplicity-DA [Alva-Manchego+ 2021]
    - Consisted 600 English sentence pairs.
    - Targeted models based on neural machine translation.
    - Manually evaluated on a DA scale of 0-100 for 3 aspects: fluency, meaning preservation, and simplicity.

[Štajner+ 2016] Sanja Štajner et al., **Shared Task on Quality Assessment for Text Simplification**, QATS, 2016
[Alva-Manchego+ 2021] Fernando Alva-Manchego et al., **The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification**, CL, 2021

# Related Work : QE for TS

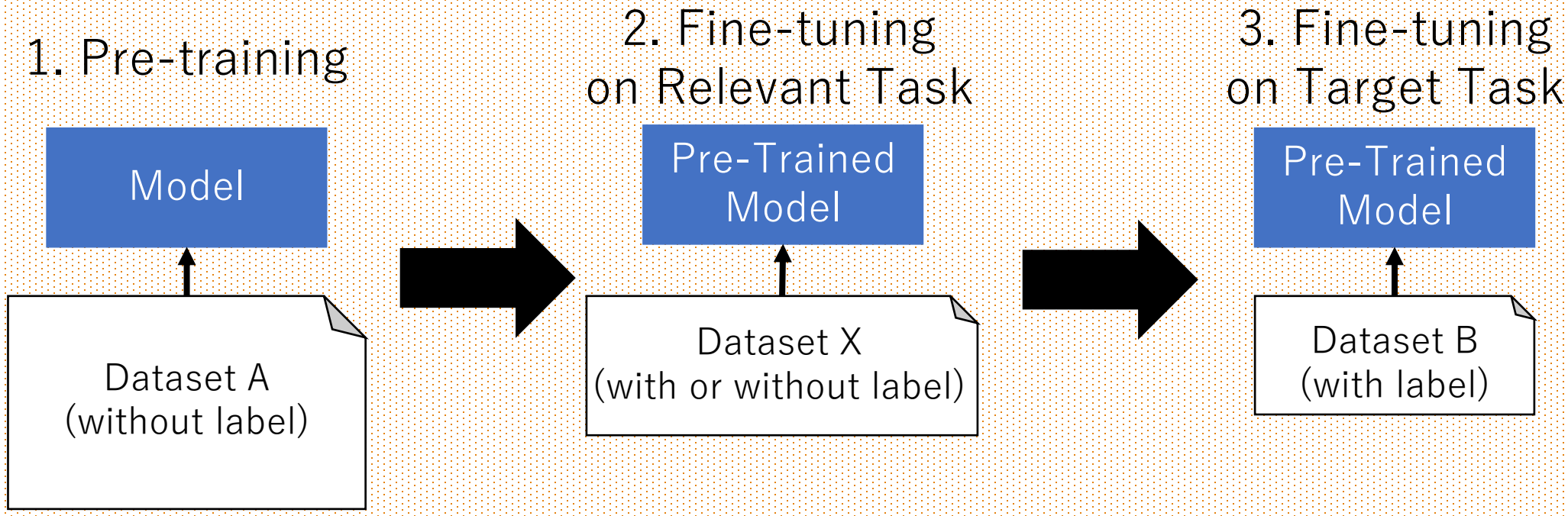
- Model
  - Kajiwara-17 [Kajiwara and Fujita 2017]
    - Word embeddings-based feature extraction
    - Trained a 3-class classification model (Good, OK, Bad) using SVM.
  - Martin-18 [Martin+ 2018]
    - Feature extraction based on machine translation evaluation metrics such as BLEU and readability metrics such as FKGL
    - Trained regression and classification models using SVM.

Issue : Difficult to train QE models based on deep learning due to the small-scale datasets.

[Kajiwaraand Fujita 2017] Tomoyuki Kajiwara Atsushi Fujita, **Semantic Features Based on Word Alignments for Estimating Quality of Text Simplification**, IJCNLP, 2017
[Martin+ 2018] Louis Martin et al., **Reference-less Quality Estimation of Text Simplification Systems**, INLG, 2018

# Related Work : Transfer Fine-tuning

The performance of transfer learning can be further improved by training on a task with similar characteristics to the target task before fine-tuning, which is called transfer fine-tuning.

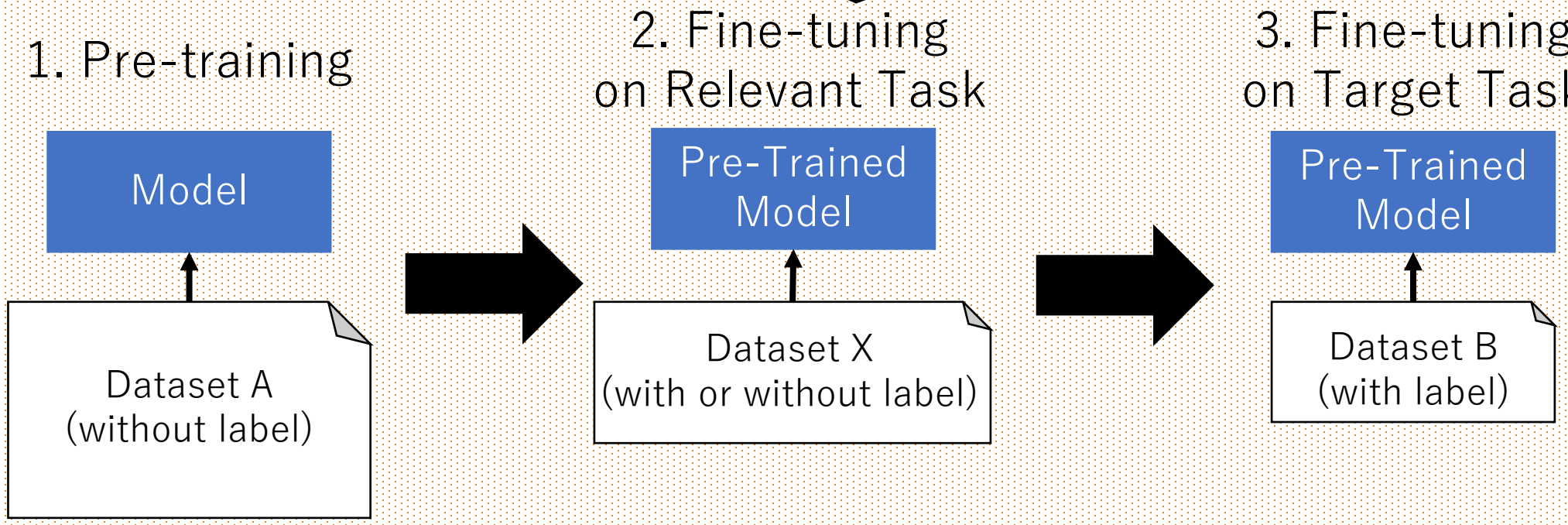



[Arase and Tsujii 2019] Yuki Arase and Jun'ichi Tsujii, **Transfer Fine-Tuning: A BERT Case Study**, EMNLP-IJCNLP, 2019
[Zhang+ 2020] Jingqing Zhang et al., **PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization**, PMLR, 2020
[Kajiwara+ 2020] Tomoyuki Kajiwara et al., **Monolingual Transfer Learning via Bilingual Translators for Style-Sensitive Paraphrase Generation**, AAAI, 2020

# Related Work : Transfer Fine-tuning

The performance of trans
on a task with similar cha
which is called transfer fine-tuning.

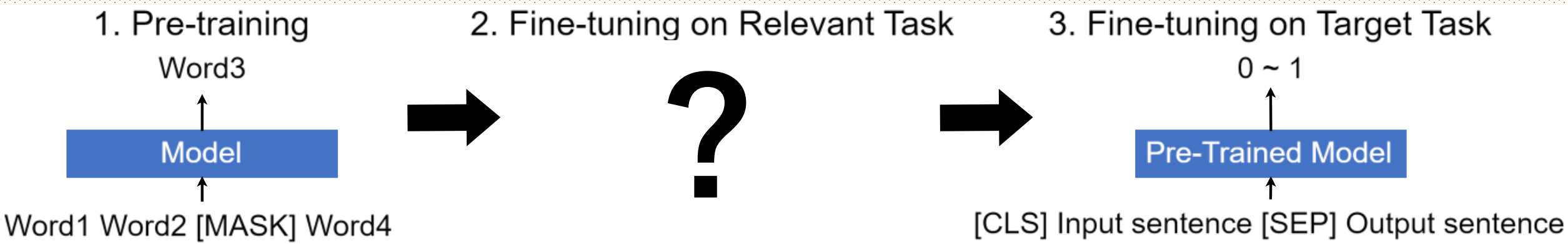
- Sentence Similarity Estimation [Arase and Tsujii 2019]
- Summarization [Zhang+ 2020]
- Paraphrase Generation [Kajiwara+ 2020]




[Arase and Tsujii 2019] Yuki Arase and Jun'ichi Tsujii, **Transfer Fine-Tuning: A BERT Case Study**, EMNLP-IJCNLP, 2019
[Zhang+ 2020] Jingqing Zhang et al., **PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization**, PMLR, 2020
[Kajiwara+ 2020] Tomoyuki Kajiwara et al., **Monolingual Transfer Learning via Bilingual Translators for Style-Sensitive Paraphrase Generation**, AAAI, 2020

# Contribution of our study

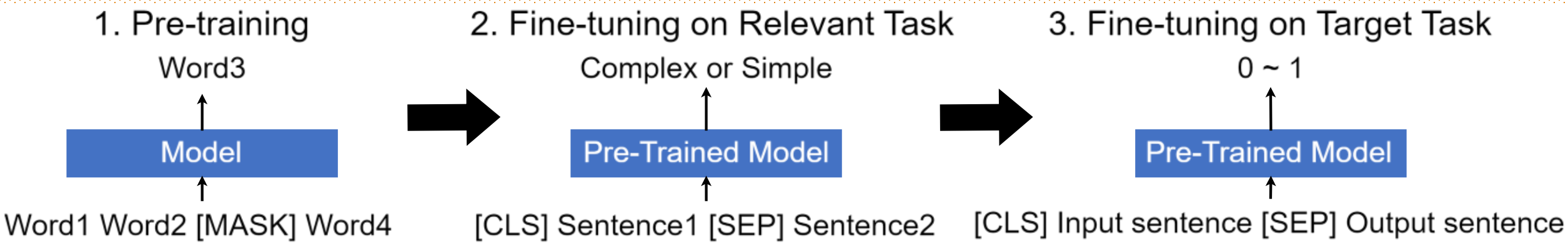
- Improved QE performance on small labeled datasets.
- Proposed additional training in QE for TS.
  - Task of identifying complex and simple sentences.
  - Using an existing large-scale parallel corpus for TS.



# Proposed Method

# Overview of the proposed method

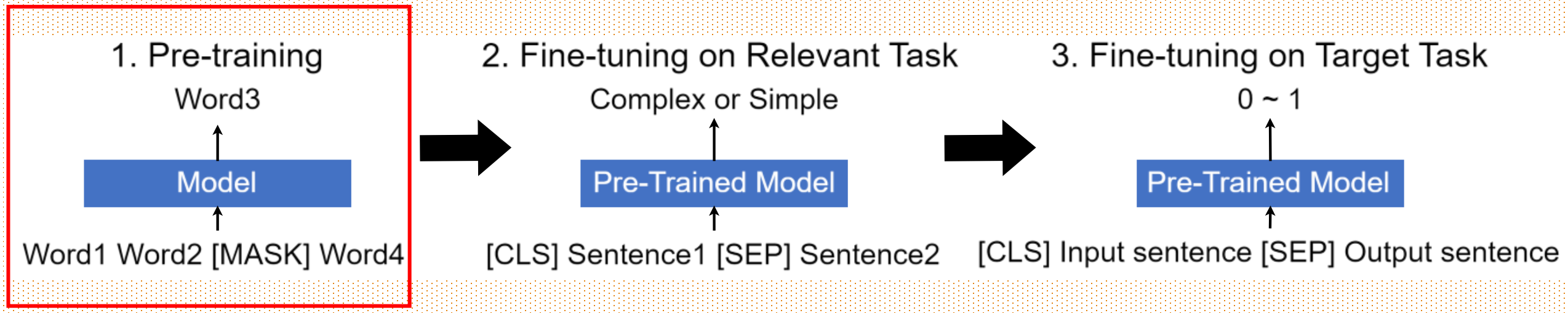
1. Pre-training : Masked Language Modeling (MLM)
2. Fine-tuning on relevant task : pseudo-QE
3. Fine-tuning on target task : QE



# 1. Pre-training : Masked Language Modeling (MLM)

Task to input masked words in a sentence and predict the masked words from the unmasked parts of the sentence.

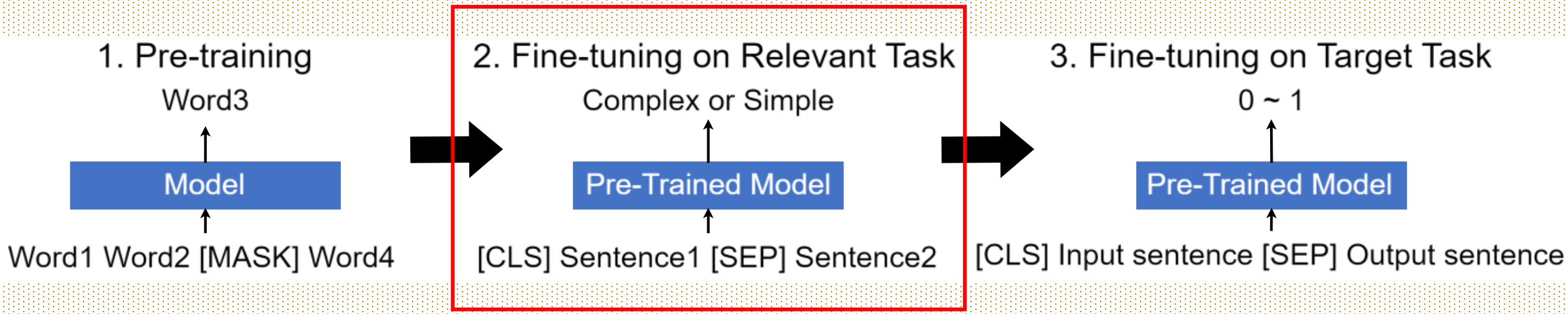
- Gain generic knowledge.
- Use pretrained models such as BERT in this study.



# 2. Fine-tuning on relevant task : pseudo-QE

Train a binary classification of whether the latter sentence is more complex or simpler.

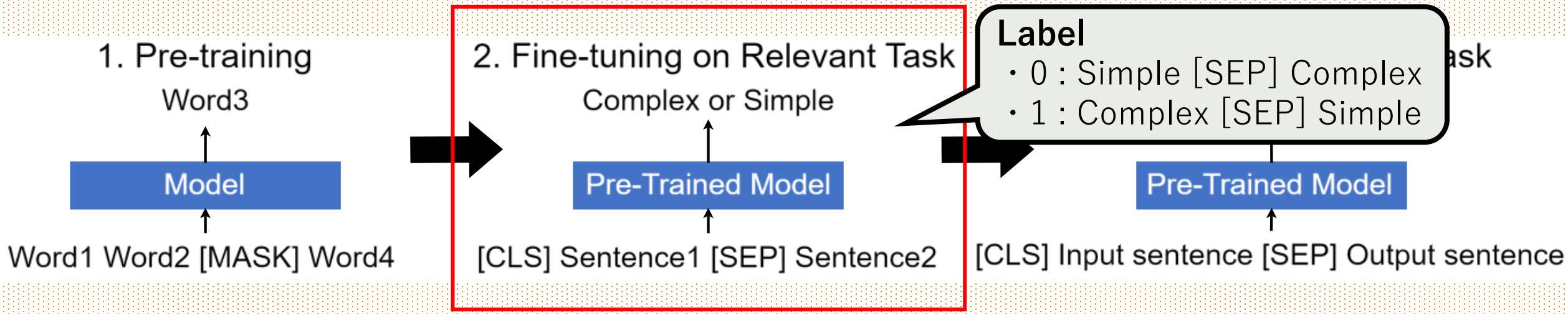
- It can be expected to improve the performance of QE for simplicity of text simplification.
- It can be used an existing large-scale parallel corpus for TS.



# 2. Fine-tuning on relevant task : pseudo-QE

Train a binary classification of whether the latter sentence is more complex or simpler.

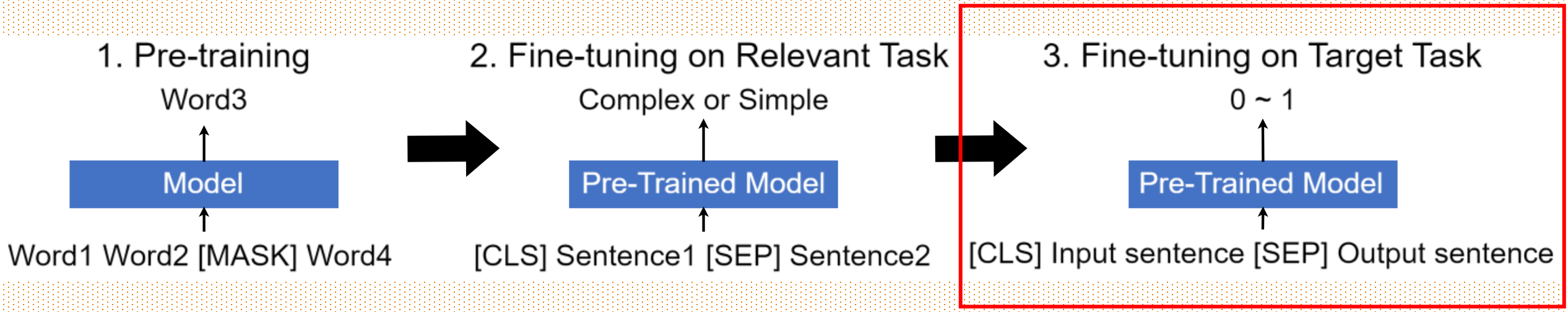
- It can be expected to improve the performance of QE for simplicity of text simplification.
- It can be used an existing large-scale parallel corpus for TS.



# 3. Fine-tuning on target task : QE

Estimate the quality of the output sentences from the input and output sentence pairs.

- Expect to be able to evaluate at a **coarse level** with pseudo-QE.
- Expect to be able to evaluate at a **finer level** with QE.



# Experiment

# Setting

- Dataset：
  - pseudo-QE：Wiki-Auto, TurkCorpus, Newsela-Auto
  - QE：Simplicity-DA
- Model：
  - Kajiwara-17
  - Martin-18
  - BERT, RoBERTa, DeBERTa
  - ＋pseudo-QE (Wikipedia)
  - ＋pseudo-QE (Newsela)
- Evaluation : Pearson correlation

| | Train | Dev | Test |
|---|---|---|---|
| Wiki-Auto | 488,332 | - | - |
| Turk Corpus | - | 2,000 | 339 |
| Newsela-Auto | 394,300 | 43,317 | 44,067 |
| Simplicity-DA | 400 | 100 | 100 |

# Result

- Improved performance on simplicity by training pseudo-QE.
- Performance on fluency of DeBERTa and meaning preservation of BERT and DeBERTa were also improved.

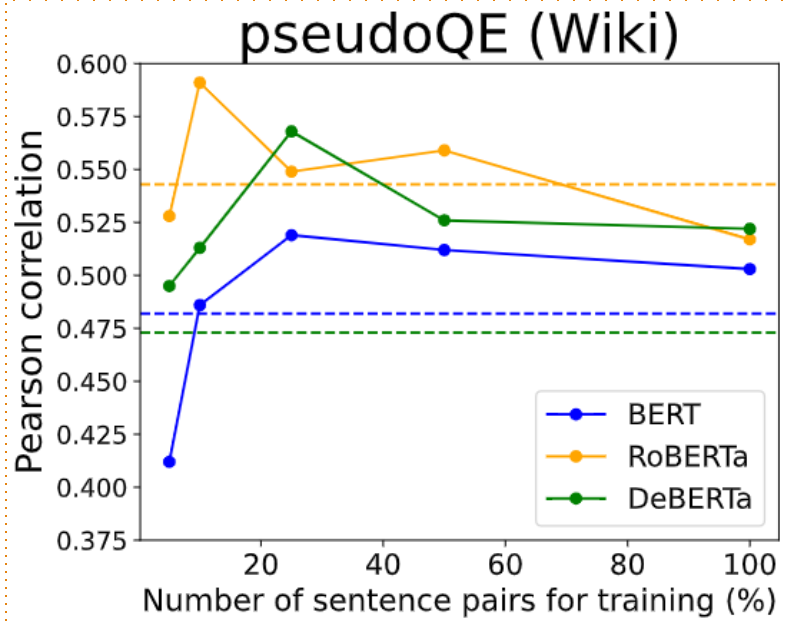
| | Fluency | Meaning | Simplicity |
|---|---|---|---|
| Kajiwara-17 | 0.405 | 0.670 | 0.373 |
| Martin-18 | 0.462 | 0.680 | 0.320 |
| BERT | 0.766 | 0.638 | 0.482 |
| + pseudo-QE (Wikipedia) | 0.739 | **0.710** | **0.503** |
| + pseudo-QE (Newsela) | 0.679 | **0.734** | 0.470 |
| RoBERTa | 0.790 | 0.779 | 0.543 |
| + pseudo-QE (Wikipedia) | 0.741 | 0.738 | 0.517 |
| + pseudo-QE (Newsela) | 0.746 | 0.764 | **0.568** |
| DeBERTa | 0.716 | 0.734 | 0.473 |
| + pseudo-QE (Wikipedia) | **0.754** | 0.728 | **0.522** |
| + pseudo-QE (Newsela) | 0.682 | **0.766** | **0.519** |

# Analysis

The change in QE performance when the amount of training data for the pseudo-QE task.

The impact of the proposed method peaks at 50,000 to 100,000 sentence pairs of training.

→There is no need to prepare a large-scale parallel corpus for text simplification with more than 100,000 sentence pairs.




Our method may be applicable to QE of text simplification in other languages, such as Italian and Japanese.

# Conclusion

- Proposed additional training in QE for TS.
  - Task of identifying complex and simple sentences.
  - Using an existing large-scale parallel corpus for TS.
  - Improved QE performance on small labeled datasets.
- Future work : focus on fluency and meaning preservation.