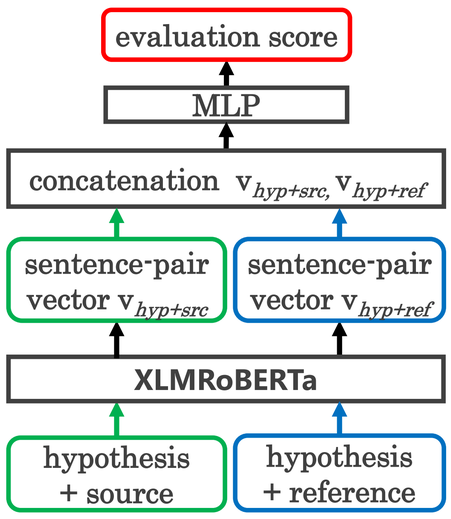
Multilingual Machine Translation Evaluation Metrics Fine-tuned on Pseudo-Negative Examples for WMT 2021 Metrics Task

Kosuke Takahashi[1], Yoichi Ishibashi [1], Katsuhito Sudoh [1,2], Satoshi Nakamura [1]
[1]Nara Institute of Science and Technology (NAIST), [2]PRESTO, Japan Science and Technology Agency
{takahashi.kosuke.th0, ishibashi.yoichi.ir3, sudoh, s-nakamura}@is.naist.jp

Abstract

- Previous studies (Takahashi et al., 2020 and Sudoh et al., 2021) and empirical experiments show that BERT-family model based metrics suffer from evaluating low quality translations.
- We prepared a pseudo-negative corpus for fine-tuning a metric model beforehand by transferring words' attributes into reversed ones.
- Experiments on the development set showed that models trained on WMT15-17/WMT18-20 and the pseudo negatives performed better than the plain ones.

Methodology



C-SPEC
Model Architecture

Training Procedure

Evaluation Model → (Fine-tuning on WMT15-16's DA) → Fine-tuned Model with Initialized MLP → (Fine-tuning on WMT15-17's DA, WMT18-20's DA, or WMT15-20's DA) → Fine-tuned Model with Initialized MLP → (Fine-tuning on WMT20's Segment-Level MQM) → C-SPEC

Fine-tuned Model with Initialized MLP → (Classification Fine-tuning on Pseudo-Negative Examples) → Fine-tuned Model with Initialized MLP → (Fine-tuning on WMT20's Segment-Level MQM) → C-SPECpn

Making of Pseudo-Negative Examples

- Apply reflection-based word attribute transfer (Ishibashi et al., 2020) to nouns of the same corpus used to fine-tune beforehand.
- Only nouns with an attribute of sex (male/female) and antonym are transferred to the target nouns with the reversed attribute.

Make 3 Categories of Hypothesis for Classification

- Original Hypothesis + Reference
- Reference + Reference
- Attribute Transferred Hypothesis + Reference

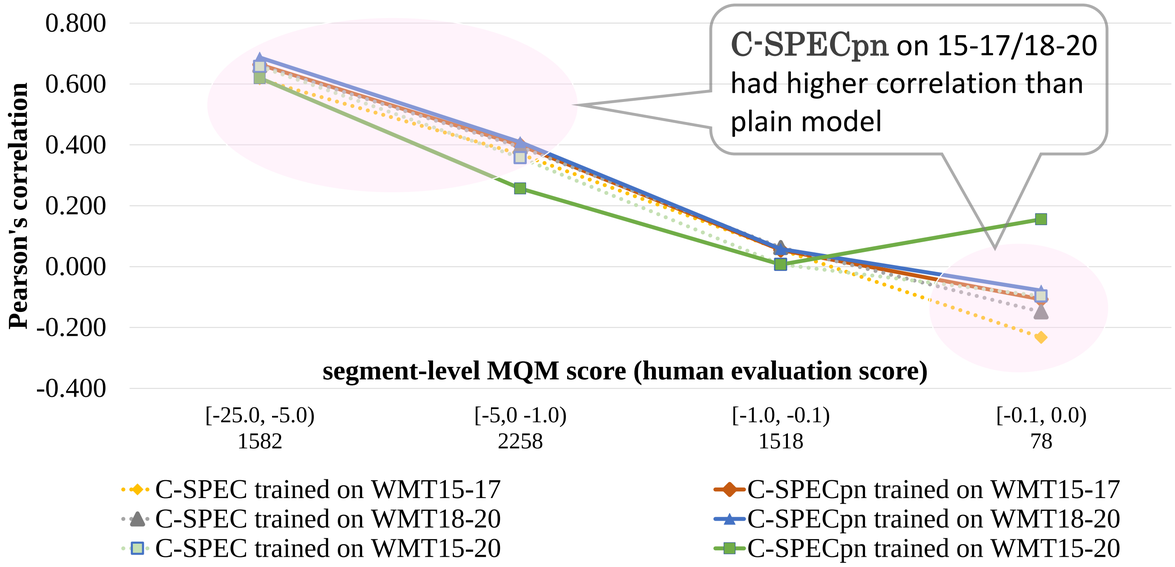
Example of Attribute Transfer

"It is our duty to remain at his sides", he said, to applause. —(Transfer Words w/ Antonym)→ "It is our duty to change at his sides", he said, to whisper.

Results on WMT20's Segment-Level MQM Human Evaluation

Pearson's correlation with MQM scores

Metric	en-de	zh-en	ave	all
C-SPEC trained on WMT15-17	0.609	0.773	0.691	0.787
C-SPEC trained on WMT18-20	0.612	0.805	0.708	0.813
C-SPEC trained on WMT15-20	0.603	0.798	0.700	0.808
C-SPECpn trained on WMT15-17	**0.626**	0.809	0.717	0.817
C-SPECpn trained on WMT18-20	0.619	**0.824**	**0.721**	**0.829**
C-SPECpn trained on WMT15-20	0.309	0.715	0.512	0.724



Pearson's correlation for each small MQM score range

- **C-SPECpn of WMT15-17 and WMT18-20 overcame the plain models.**
- Among the models, the best score was archived by C-SPECpn trained on WMT18-20.
- C-SPECpn of WMT15-17 and WMT18-20 performed better in the MQM range of [-25.0, -5.0) and [-0.1, 0.0].

This work is supported by JST PRESTO (JP-MJPR1856).