単語属性変換で作成した疑似負例データを用いた自動機械翻訳評価

○ 高橋洸丞 (NAIST)
石橋陽一 (NAIST)
須藤克仁 (NAIST/JST)
中村哲 (NAIST)

BERT[Devlin +, 2018.]を用いた自動評価手法

機械翻訳の共有タスク WMT2018/2020で
人手評価との高い相関を記録した評価手法

- **BLEURT** [Sellam +, BLEURT: Learning robust metrics for text generation, ACL 2020.]
- **C-SPEC** [Takahashi +, Automatic machine translation evaluation using source language inputs and cross-lingual language model, ACL 2020.]
- **COMET** [Rei +, COMET: A neural framework for MT evaluation, EMNLP 2020.]

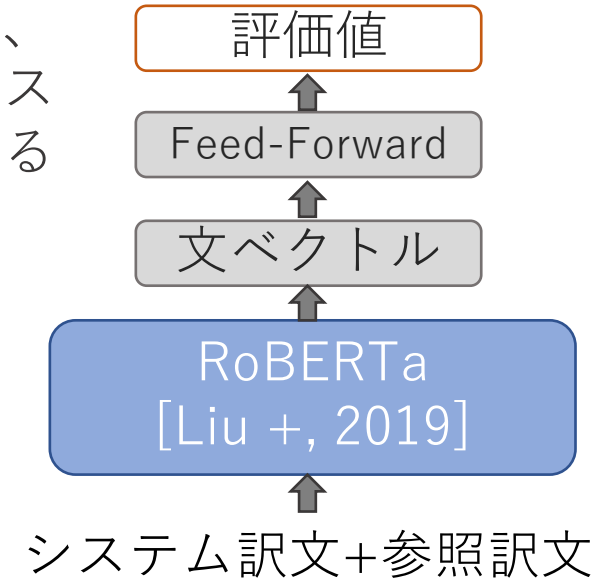
人手評価付きのデータでfine-tuningする必要がある

ノンパラメトリックな手法(BLEU[Papineni+, 2002.]やBERTscore[Zhang +, 2019.])
よりも人手評価との相関が高い

BLEURT・C-SPECの違い

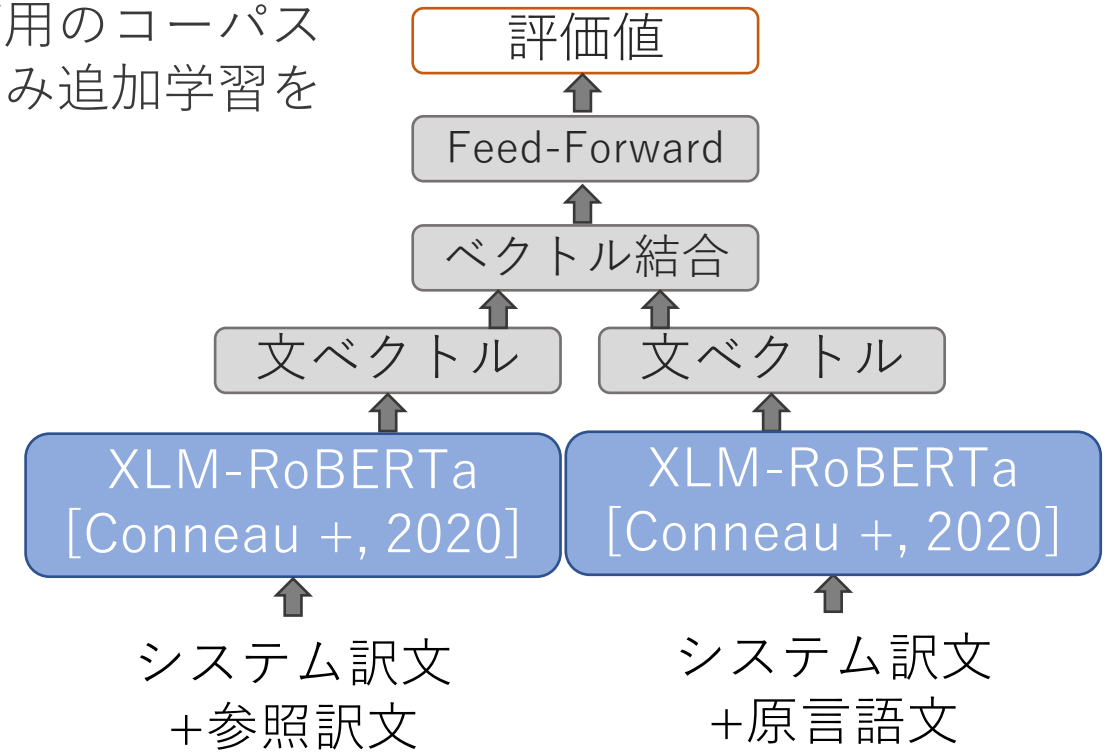
BLEURT

様々な正負の生成データで学習後、評価用のコーパスで追加学習をする



C-SPEC

評価用のコーパスでのみ追加学習をする



BLEURT・C-SPECの評価誤り

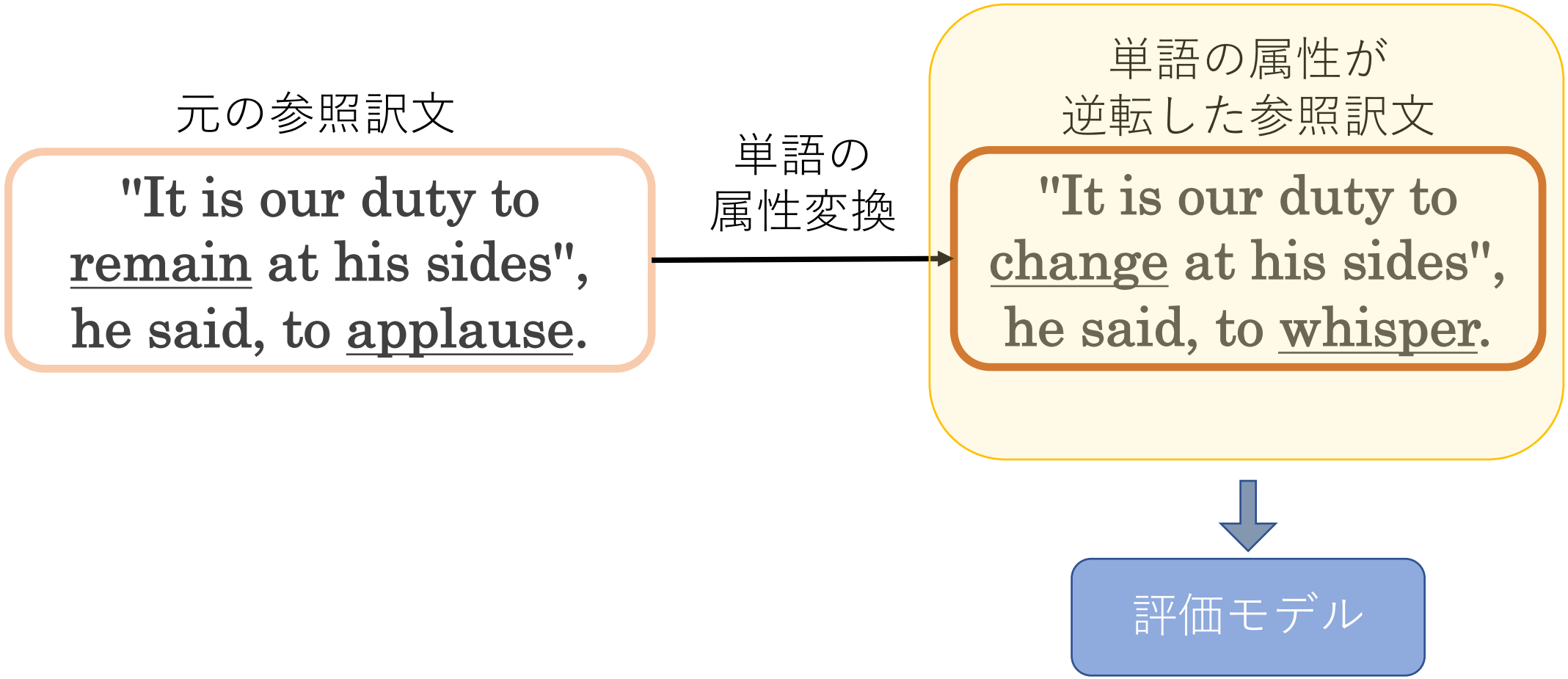
WMT17のmetrics shared taskでピアソンの相関係数を大きく低下させた文例

参照訳文	システム訳文	人手評価	BLEURT20	C-SPEC
They want to compete with delivery services.	They want to compete with airlines.	-0.861	-0.344	0.595
I'm glad we got some of these dirty licks caught on tape.	I'm glad we lick some of the dirty things they made.	-1.548	-0.088	-0.405

名詞(固有表現)の誤翻訳由来の評価誤り

述語構造の誤翻訳由来の評価誤り

提案手法 (C-SPECpn：C-SPEC fine-tuned on pseudo-negatives)
疑似的な名詞誤りを含むデータで評価モデルの追加学習



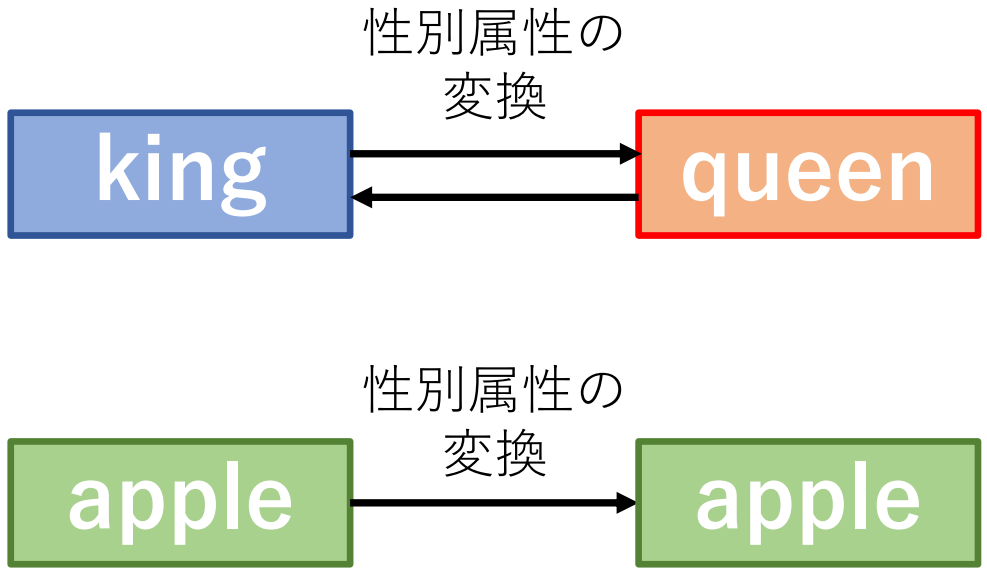
疑似負例データの作成

鏡面変換に基づく単語属性変換

[Ishibashi +, Reflection-based word attribute transfer, 2020]を参照訳文に適応

変換属性

- 性別
- 対義関係



対象の属性を
持たない単語は
変換されない

分類タスクとして負例データで学習

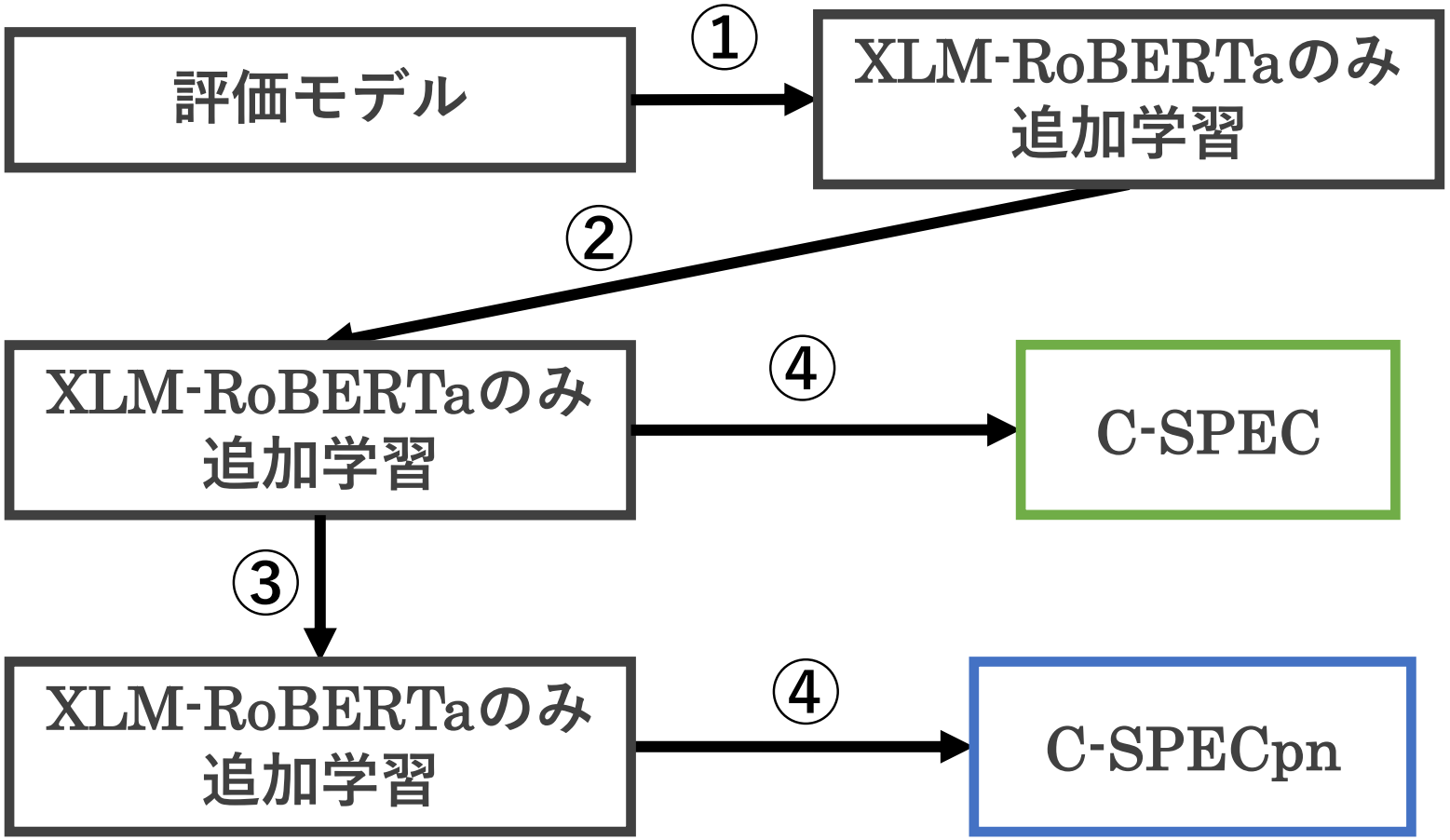
評価モデルへの入力

1. システム訳文+原言語文、システム訳文+参照訳文
   (変化なし)

2. 参照訳文+原言語文、参照訳文+参照訳文
   (参照訳文に置き換え)

3. 負例文+原言語文、負例文+参照訳文
   (負例文に置き換え)

コーパスごとの追加学習



①: WMT15-16で訓練
②: WMT18-20で訓練
③: **疑似負例データで訓練**
④: WMT20のMQMデータで訓練

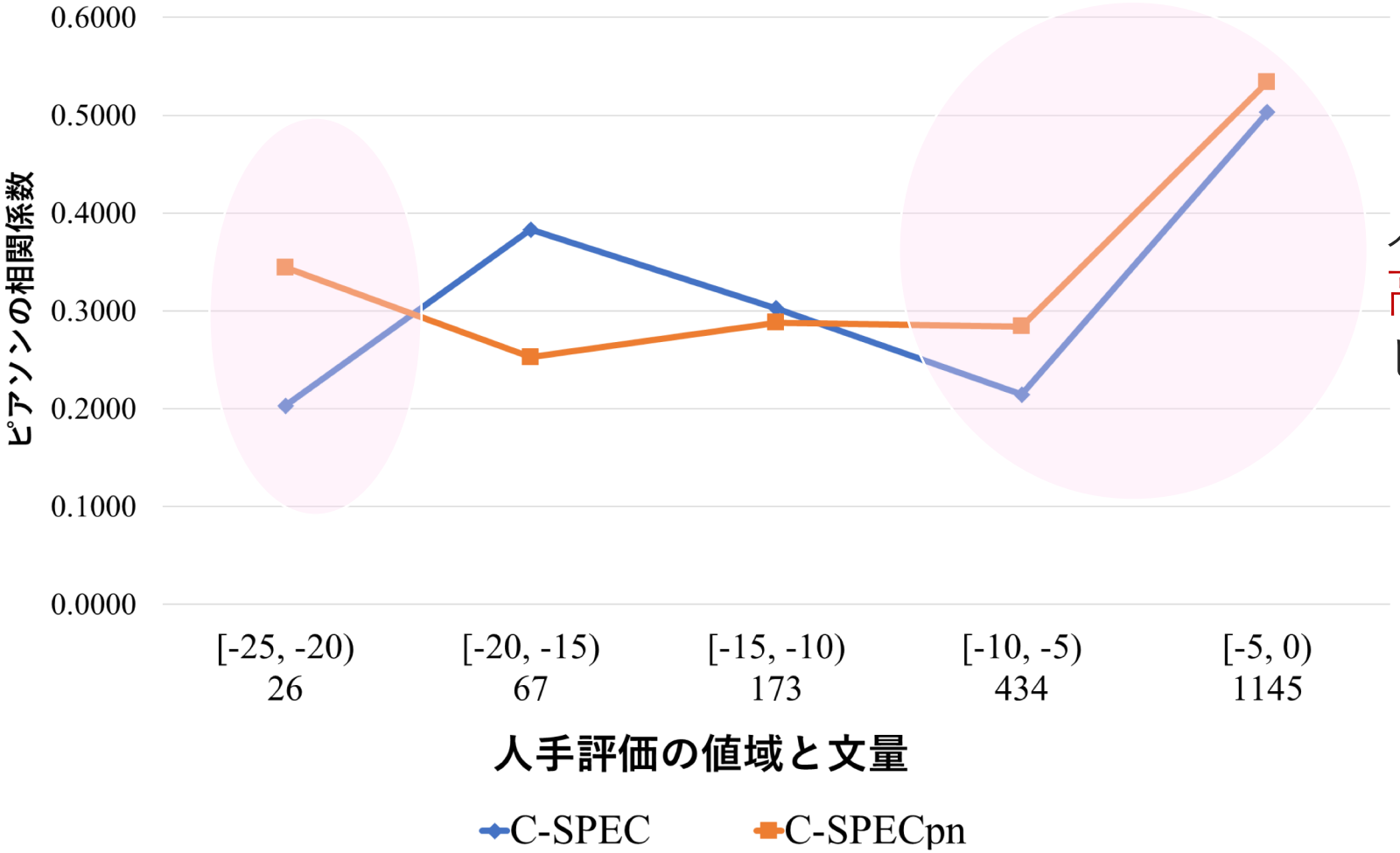
WMT20のMQMデータでの実験結果

評価モデル	英語->ドイツ語	中国語->英語	2言語対の平均	全ての言語対
ベースライン: C-SPEC	0.612	0.805	0.708	0.813
提案モデル: C-SPECpn	**0.619**	**0.824**	**0.721**	**0.829**

-NOTE-
人手評価とのピアソンの相関係数で計測
1.0に数値が近い程、人手評価との相関が高い

負例データで追加訓練することで、より高い相関が得られた

システム訳文の品質による ピアソンの相関係数の差



人手評価が低い値域と
高い値域において
ピアソンの相関係数の向上

WMT21 metrics shared taskの結果

提案モデル

Metric	Total "wins"	Language Pair en→de	en→ru	zh→en	Granularity sys	seg	Data condition news w/o HT	news w/ HT	TED
C-SPECpn	11	4	3	4	6	5	3	5	3
bleurt-20	10	4	5	1	4	6	4	3	3
COMET-MQM_2021	10	3	3	4	3	7	3	2	5
tgt-regEMT	4	1	1	2	3	1	2	1	1
COMET-QE-MQM_2021	3	1	1	1	3			3	
OpenKiwi-MQM	3	2		1	3		1	2	
RoBLEURT*	3			3	1	2	1		2
cushLEPOR(LM)	2	1		1	2		1		1
BERTScore	2	1	1		2		1		1
Prism	2		2		2		1		1
YiSi-1	2		2		2		1		1
MEE2	2	2			2		1		1
BLEU	1	1			1		1		
hLEPOR	1		1		1				1
MTEQA*	1			1	1				1
TER	1			1	1				1
chrF	1			1	1				1

Freitag +, Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain, WMT 2021 表12より引用

WMT21のチャレンジセットの事例分析

チャレンジセットでは、
①誤った参照訳文、②ドイツ語の訳出誤り　が含まれている

	原言語文	参照訳文	システム訳文	C-SPEC	C-SPECpn
①	希望水早点退吧	We hope that the flooding ends soon.	I hope the water will return early.	-1.5684	-0.6392
②	Ihr konntet denken.	You were able to think.	your konntet thinking.	-8.2422	-1.4688

C-SPECpnは参照訳文よりも原言語文との類似度を重要視している

まとめ

- 単語属性変換を用いた負例データを作成

- 負例データによる追加訓練で評価モデルの
  人手評価との相関が向上
  - 特に、高品質な翻訳文への評価に対してピアソンの相関向上

- 提案モデルのC-SPECpnは、原言語文との類似度に
  より重きをおいた評価をしている