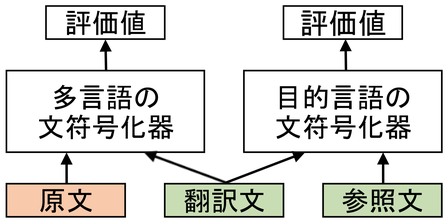
P4-33 事前学習された多言語の文符号化器を用いた機械翻訳の品質推定

嶋中宏希† 梶原智之†‡ 小町守† ShimanakaHiroki@gmail.com
†: 首都大学東京 ‡：大阪大学

1. 研究の背景と概要

- 参照文に基づく自動評価では、参照文が存在しない翻訳文の品質を評価できない
- WMTの品質推定（参照文不要の自動評価）タスクで高い性能を示す多くの手法が、大規模な対訳コーパスを用いている
- 本研究では、**多言語の文符号化器を用いた対訳コーパス不要の品質推定手法**を提案する
- 実験の結果、提案手法が多くの言語対で既存の品質推定手法の性能を上回った
- 分析の結果、他言語のデータも**言語横断的**に利用することで性能を改善でき、**zero-shot** の設定でも良好な結果が得られることが明らかになった



(a) 品質推定　(b) 自動評価

2. 関連研究

◆機械翻訳の品質推定手法

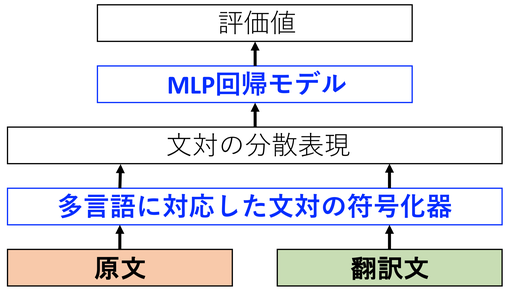
- Predictor-Estimator [Kim et al., 2017]
  - ➢ 対訳コーパス上で事前学習された Predictor と、Predictor により得られる素性から人手評価値を推定する Estimator で構成される教師あり手法
- LASER [Artetxe and Schwenk., 2019]
  - ➢ 複数言語の符号化器である LASER により得られる、原文と翻訳文のベクトル間の余弦類似度により評価する教師なし手法

両手法ともに大規模な対訳コーパスが必要

3. 提案手法

◆**多言語の文符号化器を用いた機械翻訳の品質推定手法**を提案

- 我々の先行研究であるBERTを用いた機械翻訳自動評価から以下の3点を変更し、参照文を用いない自動評価を可能にする
  - ➢ 多言語の大規模な生コーパス上で事前学習された**多言語BERT** を用いる
  - ➢ 翻訳文と参照文ではなく、**原文と翻訳文**の文対を用いて翻訳品質を推定する
  - ➢ 再学習の際には、評価対象の言語対だけでなく利用可能な**全言語対**の人手評価値付きデータを用いる



4. 実験設定

◆**人手評価値付きの学習データセット**

- 原文、参照文および翻訳文に対して人手評価値が付与
- 本研究では、原文と翻訳文と人手評価値のみを用いた
- WMT15, 16 (6,420文) の9割を学習用，1割を開発用に分割し、WMT17の各言語対 560文に対して評価

◆**文符号化器 (BERTmulti)**

- 多言語のWikipedia上で単一の符号化器を事前学習

◆**比較手法**

- Predictor-Estimator (NewsCommentary: 約0.2M文対で事前学習)
- LASER (93言語 約23M文対で事前学習)

表: WMTの自動評価タスクにおける人手評価値付きの文対数

	cs-en	de-en	fi-en	lv-en	ro-en	ru-en	tr-en	zh-en	en-ru
WMT15	500	500	500	-	-	500	-	-	500
WMT16	560	560	560	-	560	560	560	-	560
WMT17	560	560	560	560	-	560	560	560	560

5. 実験結果

◆人手評価値とのピアソンの相関係数で各手法をメタ評価

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	en-ru	Avg.
参照文なしの品質推定：									
Predictor-Estimator	0.337	0.163	-	-	0.272	-	-	0.441	0.303
LASER	0.361	0.404	0.463	0.464	0.351	0.451	**0.482**	0.352	0.416
$BERT_{multi}$	**0.548**	**0.506**	**0.695**	**0.693**	**0.592**	**0.643**	0.460	**0.648**	**0.598**
参照文に基づく自動評価：									
SentBLEU	0.435	0.432	0.571	0.393	0.484	0.538	0.512	0.468	0.479
chrF+	0.523	0.531	0.677	0.529	0.592	0.609	0.595	0.612	0.584

- 提案手法が多くの言語対で品質推定の従来手法より高い性能を示し、参照文ありの自動評価におけるベースライン手法と同等以上の性能を示す
  - ➢ **事前学習された多言語の文符号化器が機械翻訳の品質推定のために有用である**
- 学習言語が存在しない lv-en と zh-en の結果より、漢字に基づく中国語よりラテン文字に基づくラトビア語での評価のほうが高い性能を示す
  - ➢ **サブワードに基づく語彙を共有することが要因の一つと考えられる**

6. 分析

◆学習データによる性能の変化について分析するため以下の2つの設定を追加

- $BERT_{multi}$ (w/o 他言語対) ･･･ 評価対象の**言語対のみ**で学習
- $BERT_{multi}$ (Zero-shot) ･･･ 評価対象の**言語対以外**で学習

	cs-en	de-en	fi-en	ru-en	tr-en	en-ru	Avg.
全言語対	**0.548**	**0.506**	0.695	**0.592**	**0.643**	**0.648**	**0.605**
w/o 他言語対	0.474	0.442	0.638	0.424	0.533	0.599	0.518
Zero-shot	0.512	0.482	**0.697**	0.552	0.631	0.530	0.567

- w/o 他言語 (評価対象の言語対のみ) の学習より全言語のほうが性能が高い
  - ➢ **多言語の文符号化器を言語横断的に再学習することの有効性が確認できる**
- Zero-shot でも品質推定における従来手法や参照文ありの自動評価である SentBLEUよりも高い性能を示す
  - ➢ **対象言語対のためのラベル付きデータが存在しなくても、他の言語対での再学習によって高性能な品質推定を実現できると考えられる**

嶋中宏希 梶原智之 小町守, 事前学習された多言語の文符号化器を用いた機械翻訳の品質推定, 言語処理学会第26回年次大会(NLP2020), P4-33, March 18th, 2020