RUSE: Regressor Using Sentence Embeddings

Hiroki Shimanaka † Tomoyuki Kajiwara †‡ Mamoru Komachi † shimanaka-hiroki@ed.tmu.ac.jp
†: Tokyo Metropolitan University ‡: Osaka University

1. Abstract & Introduction

- Most metrics in WMT are obtained by computing based on character N-grams or word N-grams, so they can exploit only limited information for segment-level MTE. ⟶ Those models make wrong evaluations on sentences with similar meanings but different expressions.
- In this study, we propose a segment-level MTE metric using universal sentence embeddings capable of capturing semantic information that cannot be captured by surface layers.
- Our metric achieves a **state-of-the-art performance** in both segment- and system-level metrics tasks with embedding features only.

Example:

System Output: This is not a major issue.

Reference: It is nothing major.

Metrics	Score	Ranking
Human	0.892	32/560
Blend	- 0.0734	423/560
RUSE	**0.554**	**60/560**

2. RUSE: Regressor Using Sentence Embeddings

- Our metric is a regression model trained using human evaluation score with pre-trained universal sentence embeddings.

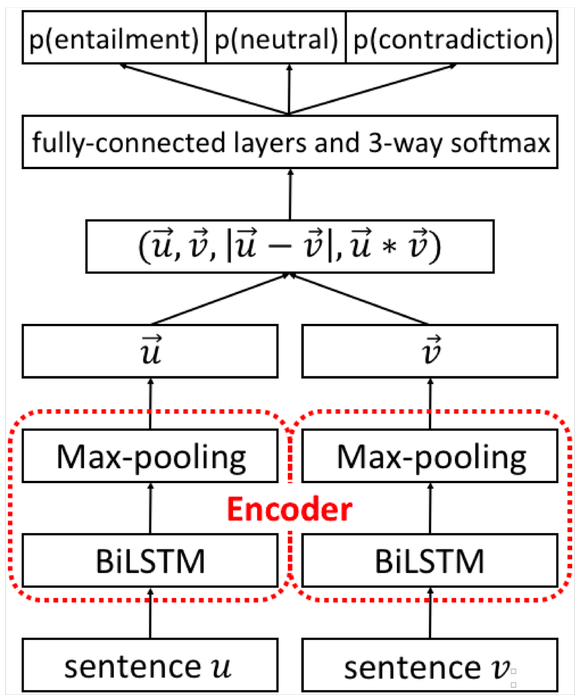

Figure1: Outline of InferSent

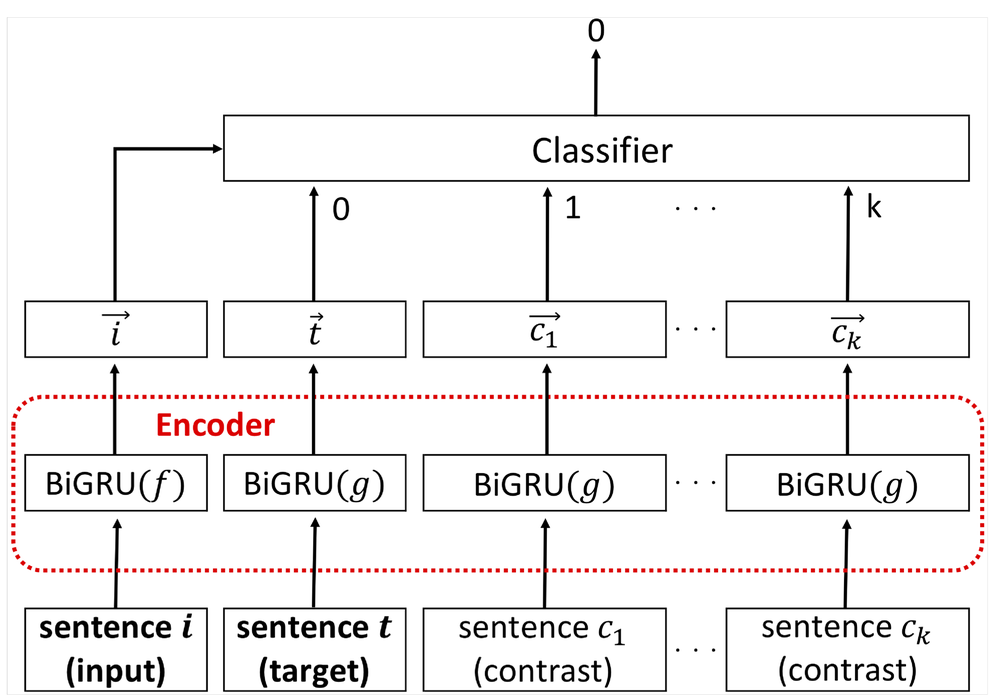

Figure2: Outline of Quick-Thought

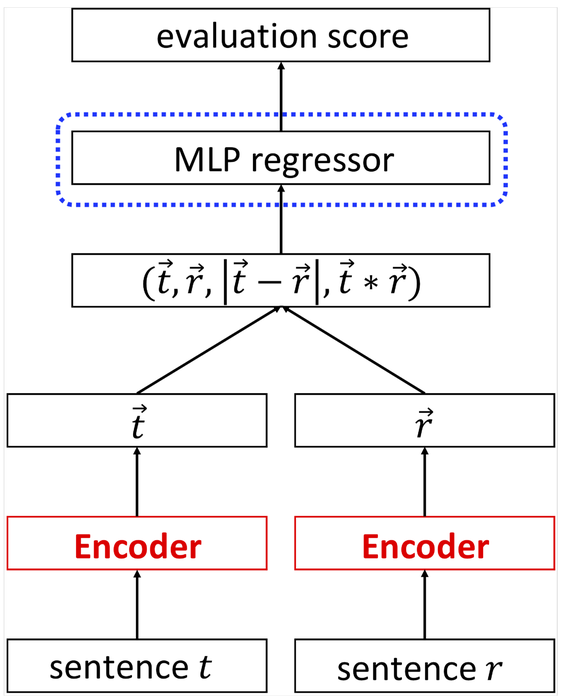

Figure3: Outline of RUSE

3. Experimental Setting

◆ Universal Sentence Embeddings

- InferSent (IS) [Conneau et al., 2017]
  - ➢ Training data: SNLI Corpus, MultiNLI Corpus
  - ➢ Dimension: 4,096
- Quick-Thought (QT) [Logeswaran and Lee, 2018]
  - ➢ Training data: Toronto-Books Corpus, UMBC corpus
  - ➢ Dimension: 4,800
- Universal Sentence Encoder (USE) [Cer et al., 2018]
  - ➢ Training data: various web sources, SNLI Corpus
  - ➢ Dimension: 512

◆ Regressor for MTE

We use MLP (Chainer) for regressor.

◆ Human evaluation data

We use DA human evaluation data in WMT.

- Train: WMT15 and WMT16 (5,360 instances)
- Development: 1/10 of train data
- Test: WMT17 (for each to-English language pairs: 560 instances)
  - ➢ We use these all DA data (9,280 instances) for submission to WMT18.

Table1: Number of human data in WMT (to-English)

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

4. Experimental Results

Table2: **Segment-level** Kendall's formulation of metric scores and DA scores in WMT18

	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en	RW
sentBLEU	0.233	0.415	0.285	0.154	0.228	0.145	0.178	0.077
BLEND [Ma et al., 2017]	**0.322**	**0.492**	0.354	0.226	**0.290**	**0.232**	**0.217**	0.434
YiSi-1 [Lo, 2018]	**0.319**	0.488	0.351	0.231	**0.300**	**0.234**	**0.211**	0.422
YiSi-1_SRL [Lo, 2018]	**0.317**	0.483	0.345	0.237	**0.306**	**0.233**	**0.209**	0.403
RUSE with IS + QT + USE	**0.347**	**0.498**	**0.368**	**0.273**	**0.311**	**0.259**	**0.218**	**0.713**

Table3: **System-level** Pearson correlation of metric scores and DA scores in WMT18

	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en	RW
BLEU [Papineni et al., 2002]	**0.970**	0.971	**0.986**	0.973	0.979	**0.657**	**0.978**	0.332
BEER [Stanojević et al ., 2015]	**0.958**	0.994	**0.985**	**0.991**	0.982	0.870	**0.976**	0.604
BLEND [Ma et al., 2017]	**0.973**	0.991	0.985	**0.994**	**0.993**	**0.801**	**0.976**	0.704
YiSi-1_SRL [Lo, 2018]	**0.965**	**0.995**	0.981	0.977	**0.992**	**0.869**	0.962	0.597
RUSE with IS + QT + USE	**0.981**	**0.997**	**0.990**	**0.991**	**0.988**	**0.853**	**0.981**	**0.832**

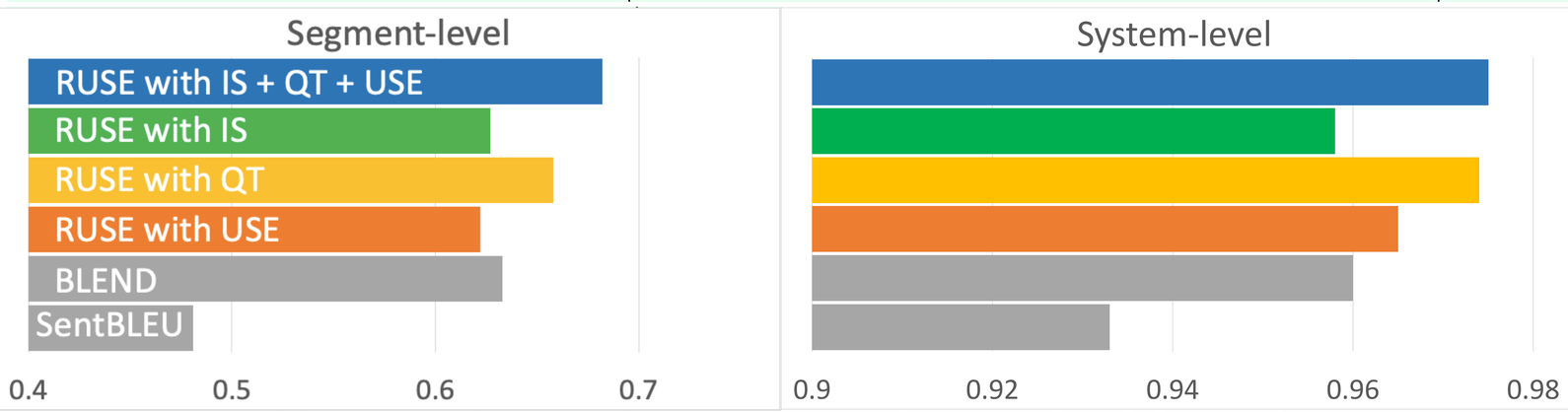

Figure4: Ablation analysis on the segment- and system-level dataset in WMT17

5. Conclusion

Our RUSE metric achieves a state-of-the-art performance in all to-English language pairs on WMT18 segment-level metrics tasks.
Based on the results of our work, we expect that the MTE metric will be further improved using these better universal sentence embeddings.
Future work is experiment using the datasets for from-English language pairs and analysis of this metric.