Metric for Automatic Machine Translation Evaluation based on Universal Sentence Representations

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

1. Abstract

- 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.
- Therefore, we propose a MTE metric by using universal sentence representations capable of capturing information that cannot be captured by local features based on character or word N-grams.
- Experimental results of the WMT-2016 dataset show that **the proposed method achieves state-of-the-art performance with sentence representation features only.**

Example

MT hypothesis: This is not a major issue.
Reference: It is nothing major.

Evaluation Metric	Score	Ranking of Scores
Human	0.892	32/560
Blend	- 0.0734	423/560
Our metric	**0.554**	**60/560**

2. Previous Works

- ReVal [Gupta et al., 2015]
  - This method is trained using datasets of sentence unit similarity scores with Tree-LSTM.
  - The training data used in this metric is small, so the learning of Tree-LSTM is unstable and accu-rate learning is difficult.
- Blend [Ma et al., 2017]
  - This method is essentially an SVR (RBF kernel) model that uses the scores of various metrics as features.
  - Features: Scores of Asiya 25 metrics and four other metrics (lexical base)

3. Proposed Method

Our metric is an SVR model trained using human evaluation scores with universal sentence representations that were trained through large-scale data.

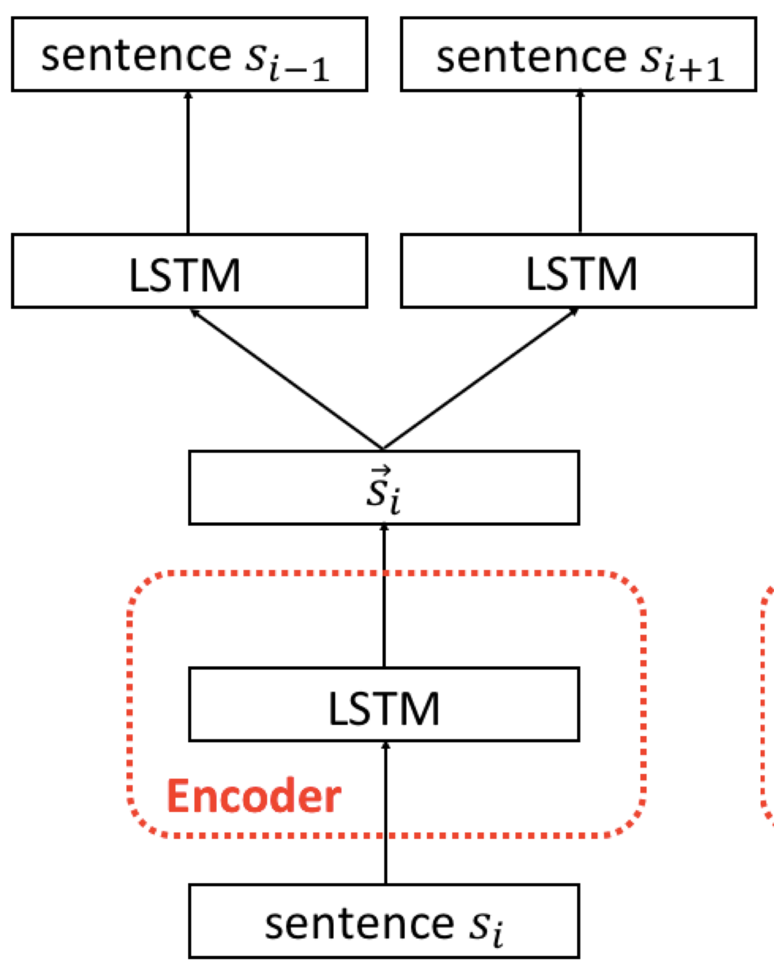

Figure 1: Outline of Skip-Thought

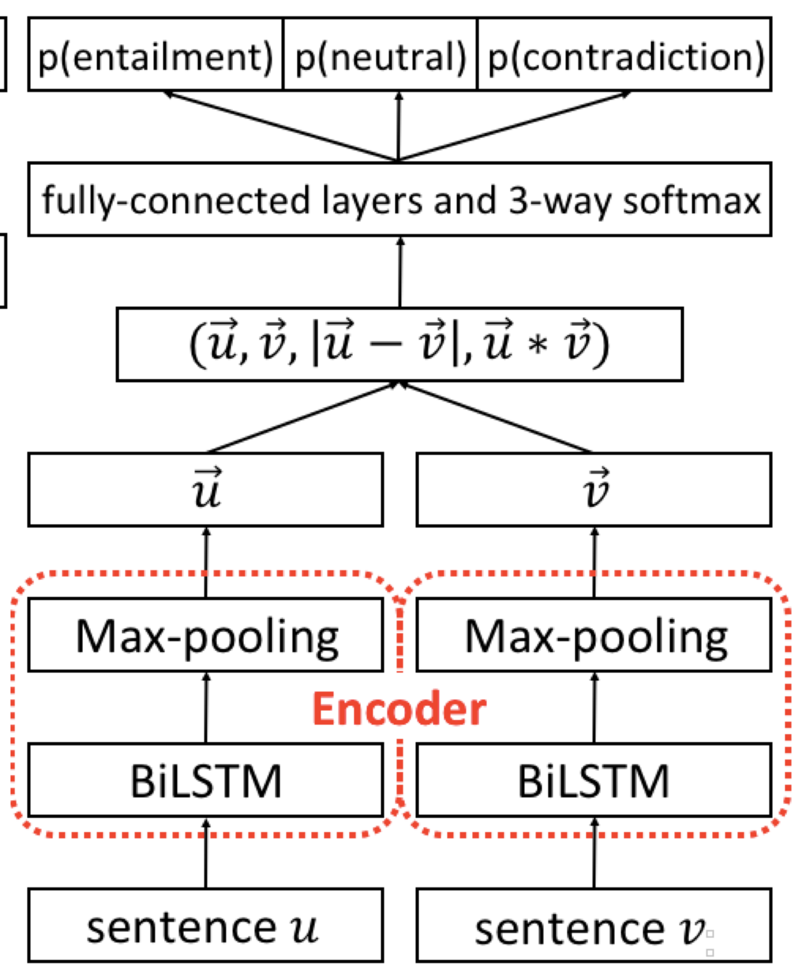

Figure 2: Outline of InferSent

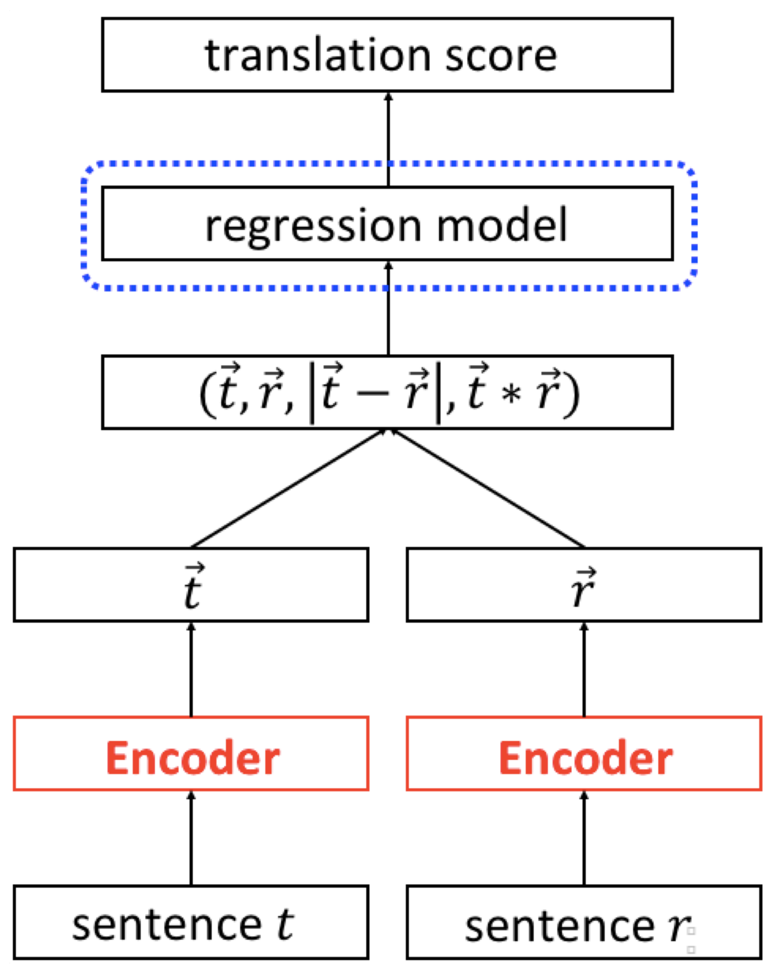

Figure 3: Outline of our metric

4. Experimental Setting

- **Universal Sentence Representations**
  We use pre-trained sentence representations that are open to the public.
- Skip-Thought [Kiros et al., 2015]
  Train Data: Toronto-Books Corpus
  Dimension: 4,800
- InferSent [Conneau et al., 2017]
  Train data: The Stanford Natural Language Inference (SNLI) Corpus
  Dimension: 4,096

- **Regression Model for MTE**
- SVR (RBF kernel) from scikit-learn
  $C \in \{0.01, 0.1, 1.0, 10\}$
  $\varepsilon \in \{0.01, 0.1, 1.0, 10\}$
  $\gamma \in \{0.01, 0.1, 1.0, 10\}$
  We performed 10-fold cross validation and grid-search.
- **Training Datasets of Human Evaluation Scores**

Table 1: Number of DA human evaluation datasets for to-English language pairs

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en
WMT-2015	500	500	500	-	500	-
WMT-2016	560	560	560	560	560	560

5. Experimental Results

Table 2: Pearson correlation of metric scores and DA human evaluations scores (newstest2016)

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	Avg.
SentBLEU	0.557	0.448	0.484	0.499	0.502	0.532	0.504
Blend [Ma et al., 2017]	0.709	0.601	0.584	0.636	0.633	0.675	0.640
$\text{DPMF}_{\text{comb}}$ [Yu et al., 2015]	**0.713**	0.584	0.598	0.627	0.615	0.663	0.633
ReVal [Gupta et al ., 2015]	0.577	0.528	0.471	0.547	0.528	0.531	0.530
SVR with Skip-Thought	0.665	0.571	0.609	**0.677**	0.608	0.599	0.622
SVR with InferSent	0.679	0.604	0.617	0.640	0.644	0.630	0.636
SVR with InferSent + Skip-Thought	0.686	**0.611**	**0.633**	0.660	**0.649**	0.646	**0.648**

6. Error Analysis

Table 3: The top 20% of MT hypotheses that were close to the meaning of the reference translations were analyzed. (Total: 672 sentence pairs)

	Only correct evaluation with Blend (Total :70 sentences)	Only correct evaluation with our metric (Total: 88 sentences)
Low word surface matching rate	26	42
Including unknown words (and short sentence length)	26 (17)	26 (2)
Other	24	31

- From the results, it is considered that our metric shows better results in MT hypotheses whose meanings are similar to those of reference translations.
- From the results about word surface matching rate, our metric can evaluate a wide range of sentence information that cannot be captured by lexical base metrics.
- From the results about unknown words, the influence of unknown words in long MT hypotheses in our metric is considered to be small.

Hiroki Shimanaka, Tomoyuki Kajiwara, Mamoru Komachi, Metric for Automatic Machine Translation Evaluation based on Universal Sentence Representations, NAACL 2018 Student Research Workshop in conjunction with NAACL HLT 2018 (NAACL-SRW 2018), June 3rd, 2018