RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation

Hiroki Shimanaka[†] **Tomoyuki Kajiwara**[†‡] **Mamoru Komachi**[†]
[†]Graduate School of Systems Design, Tokyo Metropolitan University, Tokyo, Japan
shimanaka-hiroki@ed.tmu.ac.jp, komachi@tmu.ac.jp
[‡]Institute for Datability Science, Osaka University, Osaka, Japan
kajiwara@ids.osaka-u.ac.jp

Abstract

We introduce the RUSE[1] metric for the WMT18 metrics shared task. Sentence embeddings can capture global information that cannot be captured by local features based on character or word N-grams. Although training sentence embeddings using small-scale translation datasets with manual evaluation is difficult, sentence embeddings trained from large-scale data in other tasks can improve the automatic evaluation of machine translation. We use a multi-layer perceptron regressor based on three types of sentence embeddings. The experimental results of the WMT16 and WMT17 datasets show that the RUSE metric achieves a state-of-the-art performance in both segment- and system-level metrics tasks with embedding features only.

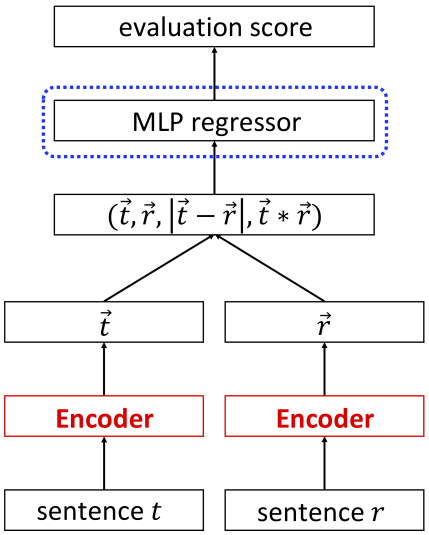

Figure 1: Outline of the RUSE metric.

1 Introduction

This study describes a segment-level metric for automatic machine translation evaluation (MTE). The MTE metrics with a high correlation with human evaluation enable the continuous integration and deployment of a machine translation (MT) system. Various MTE metrics have been proposed in the metrics task of the Workshops on Statistical Machine Translation (WMT) that was started in 2008. However, most MTE metrics are obtained by computing the similarity between an MT hypothesis and a reference based on the character or word N-grams, such as SentBLEU (Lin and Och, 2004), which is a smoothed version of BLEU (Papineni et al., 2002), Blend (Ma et al., 2017), MEANT 2.0 (Lo, 2017), and chrF++ (Popović, 2017). Therefore, they can exploit only limited information for the segment-level MTE. In other words, the MTE metrics based on character or word N-grams cannot make full use of sentence embeddings. They only check for word matches.

We extend our previous work (Shimanaka et al., 2018) and propose a segment-level MTE metric using universal sentence embeddings capable of capturing global information that cannot be captured by local features based on character or word N-grams. The experimental results in both segment- and system-level metrics tasks conducted using the datasets for to-English language pairs on WMT16 and WMT17 indicated that the proposed regression model using sentence embeddings, RUSE, achieves the best performance.

The main contributions of the study are summarized below:

- We propose a novel supervised regression model for the segment-level MTE based on universal sentence embeddings.
- We achieved a state-of-the-art performance in segment- and system-level metrics tasks on the WNT16 and WMT17 datasets for to-English language pairs without using any complex features.

[1]https://github.com/Shi-ma/RUSE

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 751–758

https://doi.org/10.18653/v1/W18-64083

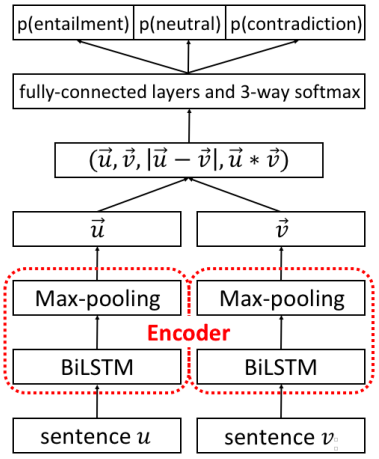

Figure 2: Outline of InferSent.

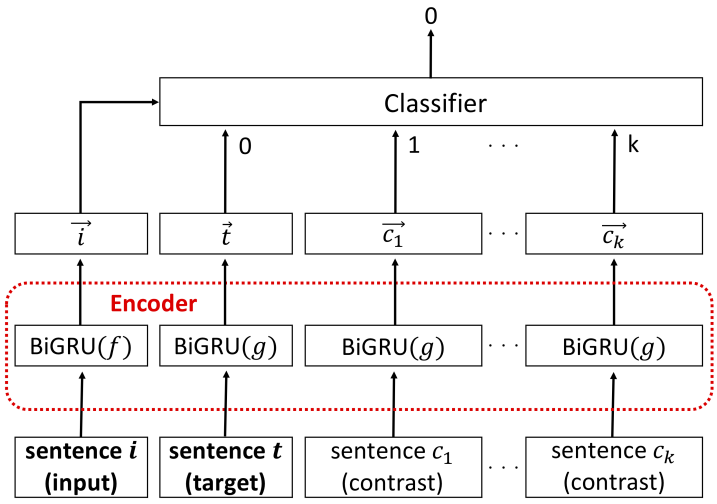

Figure 3: Outline of Quick-Thought.

2 Related Work

$\text{DPMF}_{\text{comb}}$ (Yu et al., 2015a) achieved the best performance in the WMT16 metrics task (Bojar et al., 2016). It incorporates 55 default metrics provided by the Asiya MT evaluation toolkit[2] (Giménez and Màrquez, 2010), as well as three other metrics, namely DPMF (Yu et al., 2015b), REDp (Yu et al., 2015a), and ENTFp (Yu et al., 2015a), using ranking SVM to train parameters of each metric score. DPMF evaluates the syntactic similarity between an MT hypothesis and a reference translation. REDp evaluates an MT hypothesis based on the dependency tree of the reference translation that comprises both lexical and syntactic information. ENTFp (Yu et al., 2015a) evaluates the fluency of an MT hypothesis.

After the success of $\text{DPMF}_{\text{comb}}$, Blend[3] (Ma et al., 2017) achieved the best performance in the WMT17 metrics task (Bojar et al., 2017). Similar to $\text{DPMF}_{\text{comb}}$, Blend is essentially an SVR model with RBF kernel that uses the scores of various metrics as features. It incorporates 25 lexical metrics provided by the Asiya MT evaluation toolkit, as well as four other metrics, namely BEER (Stanojević and Sima'an, 2015), CharacTER (Wang et al., 2016), DPMF, and ENTFp. BEER is a linear model based on character N-grams and replacement trees. CharacTER evaluates an MT hypothesis based on character-level edit distance.

$\text{DPMF}_{\text{comb}}$ is trained through relative ranking (RR) of human evaluation data in terms of relative ranking (RR). The quality of five MT hypotheses of the same source segment is ranked from 1 to 5 via a comparison with the reference translation. In contrast, Blend is trained through direct assessment (DA) of human evaluation data. DA provides the absolute quality scores of hypotheses by measuring to what extent a hypothesis adequately expresses the meaning of the reference translation. The experiment results in the segment-level MTE conducted using the datasets for to-English language pairs on WMT16 showed that Blend achieved a performance better than $\text{DPMF}_{\text{comb}}$. In this study, as with Blend, we propose a regression model trained using DA human evaluation data.

Instead of using local and lexical features, ReVal[4] (Gupta et al., 2015a,b) proposes using sentence-level features. It is a metric using Tree-LSTM (Tai et al., 2015) for training and capturing the holistic information of sentences. It is trained using datasets of pseudo similarity scores generated by translating RR data and out-domain datasets of similarity scores of SICK[5]. However, the training dataset used in this metric consists of approximately 21,000 sentences; thus, the learning of Tree-LSTM is unstable, and accurate learning is difficult. We use sentence embeddings trained using various RNN and Transformer as sentence information. Furthermore, we apply universal sentence embeddings to this task. These embeddings were trained using large-scale data obtained in other tasks. Therefore, the proposed approach avoids the problem of using a small dataset for training sentence embeddings.

[2] http://asiya.lsi.upc.edu/
[3] http://github.com/qingsongma/blend
[4] https://github.com/rohitguptacs/ReVal
[5] http://clic.cimec.unitn.it/composes/sick.html

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

Table 1: Number of segment-level DA human evaluation datasets for to-English language pairs[10] in WMT15 (Stanojević et al., 2015), WMT16 (Bojar et al., 2016), and WMT17 (Bojar et al., 2017).

		cs-en	de-en	fi-en	lv-en	ro-en	ru-en	tr-en	zh-en
WMT16	systems	6	10	9	-	7	10	8	-
	sentences	2,999	2,999	3,000	-	2,998	1,999	3,000	-
WMT17	systems	4	11	6	9	-	9	10	16
	sentences	3,005	3,004	3,002	2,001	-	3,001	2,017	2,001

Table 2: Number of MT systems and system-level DA human evaluation datasets for to-English language pairs in WMT16 (Bojar et al., 2016) and WMT17 (Bojar et al., 2017).

3 RUSE: Regressor Using Sentence Enbeddings

The proposed metric evaluates the MT hypothesis with universal sentence embeddings trained using large-scale data obtained in other tasks. First, we describe three types of sentence embeddings used in the proposed metric in Section 3.1. We then explain the proposed regression model and feature extraction for MTE in Section 3.2.

3.1 Universal Sentence Embeddings

Several approaches have been proposed to learn sentence embeddings. These sentence embeddings are learned through large-scale data such that they constitute potentially useful features for MTE. These have been proven effective in various NLP tasks, such as document classification and measurement of semantic textual similarity, and we call them universal sentence embeddings.

First, InferSent[6] (Conneau et al., 2017) constructs a supervised model computing universal sentence embeddings trained using Stanford Natural Language Inference (SNLI) datasets[7] (Bowman et al., 2015). The Natural Language Inference task is a classification task of sentence pairs with three labels, namely *entailment*, *contradiction*, and *neutral*; thus, InferSent can train sentence embeddings that are sensitive to differences in meaning. This model encodes a sentence pair u and v and generates features by sentence embeddings $\vec{u}$ and $\vec{v}$ with a bi-directional LSTM architecture with max pooling (Figure 2). InferSent demonstrates high performance across various document classification and semantic textual similarity tasks.

Second, Quick-Thought[8] (Logeswaran and Lee, 2018) builds an unsupervised model of universal sentence embeddings trained using some consecutive sentences. Given an input sentence and its context, a classifier distinguishes context sentences from other contrastive sentences based on their embeddings (Figure 3). For a given sentence s, its embeddings are the concatenation of the outputs of the two encoders $[f(s); g(s)]$. As a result of the training, this encoder can produce sentence embedding. Quick-Thought demonstrates high performance, especially when applied to document classification tasks.

Finally, Universal Sentence Encoder[9] (Cer et al., 2018) is trained using multitask learning, whereby a single encoding model is used to feed multiple downstream tasks. Universal Sentence Encoder supports a task to estimate the neighboring sentences for unsupervised learning and tasks conversational input–response and natural language inference for supervised learning. The unsupervised learning model trained on data drawn from a variety of web sources, such as Wikipedia, web news, web question-answer pages and discussion forums, is augmented with training

[6] https://github.com/facebookresearch/InferSent
[7] https://nlp.stanford.edu/projects/snli/
[8] https://github.com/lajanugen/S2V
[9] https://www.tensorflow.org/hub/modules/google/universal-sentence-encoder-large/2
[10] en: English, cs: Czech, de: German, fi: Finnish, ro: Romanian, ru: Russian, tr: Turkish, lv: Latvian, zh: Chinese

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	avg.
SentBLEU (Bojar et al., 2016)	0.557	0.448	0.484	0.499	0.502	0.532	0.504
COBALT-F (Bojar et al., 2016)	0.671	0.591	0.554	0.639	0.618	0.627	0.617
METRICS-F (Bojar et al., 2016)	0.696	0.601	0.557	0.662	0.618	0.649	0.631
$\text{DPMF}_{\text{comb}}$ (Bojar et al., 2016)	0.713	0.584	0.598	0.627	0.615	0.663	0.633
RUSE (MLP) with IS+QT+USE	0.717	**0.661**	**0.682**	**0.725**	0.663	0.661	**0.685**
RUSE (SVR) with IS+QT+USE	**0.720**	0.632	0.678	0.708	**0.670**	**0.675**	0.681

Table 3: Segment-level Pearson correlation of metric scores and DA human evaluation scores for to-English language pairs in WMT16. IS: InferSent; QT: Quick-Thought; and USE: Universal Sentence Encoder.

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg.
SentBLEU (Bojar et al., 2017)	0.435	0.432	0.571	0.393	0.484	0.538	0.512	0.481
chrF++ (Bojar et al., 2017)	0.523	0.534	0.678	0.520	0.588	0.614	0.593	0.579
MEANT 2.0 (Bojar et al., 2017)	0.578	0.565	0.687	0.586	0.607	0.596	0.639	0.608
Blend (Bojar et al., 2017)	0.594	0.571	0.733	0.577	0.622	0.671	0.661	0.633
RUSE (MLP) with IS+QT+USE	0.614	0.637	**0.756**	**0.705**	**0.680**	0.704	0.677	0.682
RUSE (SVR) with IS+QT+USE	**0.624**	**0.644**	0.750	0.697	0.673	**0.716**	**0.691**	**0.685**

Table 4: Segment-level Pearson correlation of metric scores and DA human evaluation scores for to-English language pairs in WMT17. IS: InferSent; QT: Quick-Thought; and USE: Universal Sentence Encoder.

on supervised data from the SNLI corpus. Universal Sentence Encoder demonstrates a higher performance across various document classification and semantic textual similarity tasks compared to InferSent.

3.2 Regression Model for MTE

This study proposes a segment-level MTE metric for to-English language pairs. This problem can be treated as a regression problem that estimates the translation quality as a real number from an MT hypothesis t and a reference translation r. Once d-dimensional sentence vectors $\vec{t}$ and $\vec{r}$ are generated, the proposed model applies the following three matching methods to extract the relations between t and r (Figure 1).

- Concatenation: $(\vec{t}, \vec{r})$
- Element-wise product: $\vec{t} * \vec{r}$
- Absolute element-wise difference: $|\vec{t} - \vec{r}|$

Thus, we perform regression using $4d$-dimensional features of $\vec{t}$, $\vec{r}$, $\vec{t} * \vec{r}$ and $|\vec{t} - \vec{r}|$.

4 Experiments

We performed experiments using the evaluation datasets of the WMT metrics task to verify the performance of the proposed metric.

4.1 Setup

Datasets. We used segment-level datasets for to-English language pairs from the WMT15 (Stanojević et al., 2015), WMT16 (Bojar et al., 2016), and WMT17 (Bojar et al., 2017) metrics tasks as summarized in Table 1. For testing, we also used system-level datasets from the WMT16 and WMT17 metrics tasks as summarized in Table 2.

Training. We divided the dataset for training and development at a 9:1 ratio. First, for testing in WMT16, we divided the segment-level dataset of WMT15 into 1800 instances for training and 200 instances for development. Next, for testing in WMT17, we divided the segment-level datasets of WMT15 and WMT16 into 4824 instances for training and 536 instances for development. Finally, for submission to WMT18, we divided the segment-level dataset of WMT15, WMT16, and WMT17 into 8352 instances for training and 928 instances for development.

Testing. We scored each sentence using our metric for to-English language pairs in both segment and system levels. For testing on the system-level metrics task, we calculated the average score for each system as a system-level score. We evaluated our metric using the Pearson correlation coefficient between the metric scores and the DA hu-

	cs-en	de-en	fi-en	ro-en	ru-en	tr-en	avg.
BLEU (Bojar et al., 2016)	0.989	0.808	0.864	0.840	0.837	0.895	0.872
BEER (Bojar et al., 2016)	0.990	0.879	0.972	0.852	0.901	0.982	0.929
MPEDA (Bojar et al., 2016)	**0.993**	0.937	0.976	0.932	0.929	0.982	0.958
ReVal (Bojar et al., 2016)	0.986	**0.985**	0.970	0.957	**0.976**	0.958	0.972
RUSE (MLP) with IS+QT+USE	0.990	0.968	**0.977**	**0.962**	0.953	**0.991**	**0.974**
RUSE (SVR) with IS+QT+USE	0.990	0.954	0.976	0.940	0.944	0.984	0.965

Table 5: System-level Pearson correlation of metric scores and DA human evaluation scores for to-English language pairs in WMT16. IS: InferSent; QT: Quick-Thought; and USE: Universal Sentence Encoder.

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg.
BLEU (Bojar et al., 2017)	0.971	0.923	0.903	0.979	0.912	0.976	0.864	0.933
UHH_TSKM (Bojar et al., 2017)	**0.996**	0.937	0.921	0.990	0.914	0.987	0.902	0.950
BEER (Bojar et al., 2017)	0.972	0.960	0.955	0.978	0.936	0.972	0.902	0.954
Blend (Bojar et al., 2017)	0.968	**0.976**	0.958	0.979	**0.964**	0.984	0.894	0.960
RUSE (MLP) with IS+QT+USE	0.995	0.964	**0.985**	**0.996**	0.956	**0.993**	**0.937**	**0.975**
RUSE (SVR) with IS+QT+USE	**0.996**	0.964	0.983	0.988	0.951	**0.993**	0.930	0.972

Table 6: System-level Pearson correlation of metric scores and DA human evaluation scores for to-English language pairs in WMT17. IS: InferSent; QT: Quick-Thought; and USE: Universal Sentence Encoder.

man scores.

Features. Publicly available pre-trained sentence embeddings, such as InferSent[6], Quick-Thought[8], and Universal Sentence Encoder[9], were used as the features mentioned in Section 3. InferSent is a collection of 4096-dimensional sentence embeddings trained on both 560,000 sentences of the SNLI dataset (Bowman et al., 2015) and 433,000 sentences of the MultiNLI dataset (Williams et al., 2018). Quick-Thought is a collection of 4800-dimensional sentence embeddings trained on both 45 million sentences of the BookCorpus dataset (Zhu et al., 2015) and 129 million sentences of the UMBC corpus (Han et al., 2013). Universal Sentence Encoder is a collection of 512-dimensional sentence embeddings trained on many sentences from a variety of web Sources, such as Wikipedia, web news, web question-answer pages, and discussion forums.

Model. Our regression model used a multi-layer perceptron (MLP) from Chainer[11] and Support Vector Regression (SVR) from sckit-learn[12] with the features mentioned in Section 3.2.

MLP regressor. Hyper-parameters were determined through grid search in the following parameters using the development data. We used ReLU as an activation function in all layers.

- Number of layers $\in \{1, 2, 3\}$
- Number of units $\in \{512, 1024, 2048, 4096\}$
- Batch size $\in \{64, 128, 256, 512, 1024\}$
- Dropout rate $\in \{0.1, 0.3, 0.5\}$
- Optimizer $\in \{\text{Adam}\}$

SVR. We used an SVR model with the RBF kernel. The hyper-parameters were determined through a 10-fold cross validation in the following parameters using the training and development data.

- $C \in \{0.1, 1.0, 10\}$
- $\epsilon \in \{0.01, 0.1, 1.0\}$
- $\gamma \in \{0.001, 0.01, 0.1\}$

Baseline Metrics. We compared the proposed metric with the four baseline metrics for each dataset. One is BLEU, which is the de facto standard metric for machine translation evaluation. The others are the top three metrics in each task.

[11] https://chainer.org/
[12] http://scikit-learn.org/

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg.
Blend (Bojar et al., 2017)	0.594	0.571	0.733	0.577	0.622	0.671	0.661	0.633
RUSE (MLP) with IS+QT+USE	**0.614**	**0.637**	**0.756**	**0.705**	**0.680**	**0.704**	**0.677**	**0.682**
RUSE (MLP) with IS	0.556	0.568	0.706	0.650	0.626	0.649	0.634	0.627
RUSE (MLP) with QT	0.601	0.587	0.737	0.685	0.661	0.692	0.647	0.658
RUSE (MLP) with USE	0.592	0.596	0.681	0.621	0.598	0.645	0.620	0.622

Table 7: Ablation analysis on the segment-level dataset in WMT17.

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	avg.
Blend (Bojar et al., 2017)	0.968	**0.976**	0.958	0.979	**0.964**	0.984	0.894	0.960
RUSE (MLP) with IS+QT+USE	0.995	0.964	0.985	0.996	0.956	**0.993**	**0.937**	**0.975**
RUSE (MLP) with IS	0.984	0.972	0.963	0.969	0.955	0.982	0.881	0.958
RUSE (MLP) with QT	0.997	0.952	**0.997**	**0.998**	0.945	0.992	0.936	0.974
RUSE (MLP) with USE	**0.999**	0.947	0.982	0.975	0.958	0.960	0.932	0.965

Table 8: Ablation analysis on the system-level dataset in WMT17.

4.2 Result

Segment-level metrics task. Tables 3 and 4 show the experimental results on the segment level. Our proposed metrics achieved the best performance in all to-English language pairs. For the segment-level tasks, both MLP and SVR regressors outperformed the state-of-the-art metrics.

System-level metrics task. Tables 5 and 6 present the experimental results on the system level. Our proposed metric based on the MLP regressor achieved the best performance in several to-English language pairs and outperformed the state-of-the-art metrics on average.

4.3 Discussion

These results indicated that adopting universal sentence embeddings in MTE is possible by training a regression model using DA human evaluation data. Blend is an ensemble method using combinations of various MTE metrics as features; hence, our results showed that universal sentence embeddings can more accurately consider the similarity between the MT hypothesis and the reference than a complex model.

MLP vs. SVR in the RUSE metric. These experimental results showed that in the RUSE metric, MLP performed better than SVR in many cases. In addition, MLP can be trained and inferred faster than SVR by making effective use of GPU. Therefore, we submitted a model of RUSE (MLP) with IS+QT+USE trained on the whole dataset to WMT18.

Ablation analysis. Tables 7 and 8 show that our metric with Quick-Thought feature only outperformed the state-of-the-art metrics in both segment- and system-level metrics tasks. Quick-Thought is an unsupervised model of universal sentence embeddings trained using some consecutive sentences. Therefore, Quick-Thought can be trained in corpora of languages other than English. Our method is effective if there are universal sentence embeddings and DA human evaluation data. Thus, our method with Quick-Thought may be effective in MTE for other than to-English language pairs.

5 Conclusions

In this study, we applied universal sentence embeddings to MTE based on the DA of human evaluation data. Our segment-level MTE metric RUSE achieved the best performance in both segment- and system-level metrics tasks on the WMT16 and WMT17 datasets. We conclude that:

- Universal sentence embeddings can more comprehensively consider information than an ensemble metric using combinations of various MTE metrics based on the features of character or word N-grams.
- Universal sentence embeddings trained on a large-scale dataset are more effective than sentence embeddings trained on a small or limited in-domain dataset.

References

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 Metrics Shared Task. In *Proceedings of the First Conference on Machine Translation*, pages 199–231.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *arXiv preprint arXiv:1803.11175v2*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Jesús Giménez and Lluís Màrquez. 2010. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-) Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015a. ReVal: A Simple and Effective Machine Translation Evaluation Metric Based on Recurrent Neural Networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015b. Machine Translation Evaluation using Recurrent Neural Networks. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 380–384.

Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 44–52. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507.

Chi-Kiu Lo. 2017. MEANT 2.0: Accurate Semantic MT Evaluation for Any Output Language. In *Proceedings of the Second Conference on Machine Translation*, pages 589–597.

Lajanugen Logeswaran and Honglak Lee. 2018. An Efficient Framework for Learning Sentence Representations. In *International Conference on Learning Representations*.

Qingsong Ma, Yvette Graham, Shugen Wang, and Qun Liu. 2017. Blend: a Novel Combined MT Metric Based on Direct Assessment —CASICT-DCU submission to WMT17 Metrics Task. In *Proceedings of the Second Conference on Machine Translation*, pages 598–603.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2017. chrF++: Words Helping Character N-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. Metric for Automatic Machine Translation Evaluation based on Universal Sentence Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 106–111.

Miloš Stanojević, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 Metrics Shared Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273.

Miloš Stanojević and Khalil Sima'an. 2015. BEER 1.1: ILLC UvA Submission to Metrics and Tuning Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 396–401.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1556–1566.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTER: Translation Edit Rate on Character Level. In *Proceedings of the First Conference on Machine Translation*, pages 505–510.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122. Association for Computational Linguistics.

Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. 2015a. CASICT-DCU Participation in WMT2015 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 417–421.

Hui Yu, Xiaofeng Wu, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2015b. An Automatic Machine Translation Evaluation Metric Based on Dependency Parsing Model. *arXiv preprint arXiv:1508.01996*.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.