# Putting Evaluation in Context: Contextual Embeddings improve Machine Translation Evaluation

**Nitika Mathur** **Timothy Baldwin** **Trevor Cohn**
School of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia

nmathur@student.unimelb.edu.au {tbaldwin,tcohn}@unimelb.edu.au

## Abstract

Accurate, automatic evaluation of machine translation is critical for system tuning, and evaluating progress in the field. We proposed a simple unsupervised metric, and additional supervised metrics which rely on contextual word embeddings to encode the translation and reference sentences. We find that these models rival or surpass all existing metrics in the WMT 2017 sentence-level and system-level tracks, and our trained model has a substantially higher correlation with human judgements than all existing metrics on the WMT 2017 to-English sentence level dataset.

## 1 Introduction

Evaluation metrics are a fundamental component of machine translation (MT) and other language generation tasks. The problem of assessing whether a translation is both adequate and coherent is a challenging text analysis problem, which is still unsolved, despite many years of effort by the research community. Shallow surface-level metrics, such as BLEU and TER (Papineni et al., 2002; Snover et al., 2006), still predominate in practice, due in part to their reasonable correlation to human judgements, and their being parameter free, making them easily portable to new languages. In contrast, trained metrics (Song and Cohn, 2011; Stanojevic and Sima'an, 2014; Ma et al., 2017; Shimanaka et al., 2018), which are learned to match human evaluation data, have been shown to result in a large boost in performance.

This paper aims to improve over existing MT evaluation methods, through developing a series of new metrics based on contextual word embeddings (Peters et al., 2018; Devlin et al., 2019), a technique which captures rich and portable representations of words in context, which have been shown to provide important signal to many other NLP tasks (Rajpurkar et al., 2018). We propose a simple untrained model that uses off-the-shelf contextual embeddings to compute approximate recall, when comparing a reference to an automatic translation, as well as trained models, including: a recurrent model over reference and translation sequences, incorporating attention; and the adaptation of an NLI method (Chen et al., 2017) to MT evaluation. These approaches, though simple in formulation, are highly effective, and rival or surpass the best approaches from WMT 2017. Moreover, we show further improvements in performance when our trained models are learned using noisy crowd-sourced data, i.e., having single annotations for more instances is better than collecting and aggregating multiple annotations for single instances. The net result is an approach that is more data efficient than existing methods, while producing substantially better human correlations.[1]

## 2 Related work

MT metrics attempt to automatically predict the quality of a translation by comparing it to a reference translation of the same source sentence.

Metrics such as BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) use $n$-gram matching or more explicit word alignment to match the system output with the reference translation. Character-level variants such as BEER, CHRF and CHARACTER overcome the problem of harshly penalising morphological variants, and perform surprisingly well despite their simplicity (Stanojevic and Sima'an, 2014; Popović, 2015; Wang et al., 2016).

In order to allow for variation in word choice and sentence structure, other metrics use information from shallow linguistic tools such as POS-taggers, lemmatizers and synonym dictionaries (Banerjee and Lavie, 2005; Snover et al., 2006; Liu et al., 2010), or deeper linguistic informa-

[1] code is available at https://github.com/nitikam/mteval-in-context



*Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808
Florence, Italy, July 28 - August 2, 2019. 

tion such as semantic roles, dependency relationships, syntactic constituents, and discourse roles (Giménez and Màrquez, 2007; Castillo and Estrella, 2012; Guzmán et al., 2014). On the flip side, it is likely that these are too permissive of mistakes.

More recently, metrics such as MEANT_2.0 (Lo, 2017) have adopted word embeddings (Mikolov et al., 2013) to capture the semantics of individual words. However, classic word embeddings are independent of word context, and context is captured instead using hand-crafted features or heuristics.

Neural metrics such as ReVal and RUSE solve this problem by directly learning embeddings of the entire translation and reference sentences. ReVal (Gupta et al., 2015) learns sentence representations of the MT output and reference translation as a Tree-LSTM, and then models their interactions using the element-wise difference and angle between the two. RUSE (Shimanaka et al., 2018) has a similar architecture, but it uses pretrained sentence representations instead of learning the sentence representations from the data.

The Natural Language Inference (NLI) task is similar to MT evaluation (Padó et al., 2009): a good translation entails the reference and vice-versa. An irrelevant/wrong translation would be neutral/contradictory compared to the reference. An additional complexity is that MT outputs are not always fluent. On the NLI datasets, systems that include pairwise word interactions when learning sentence representations have a higher accuracy than systems that process the two sentences independently (Rocktäschel et al., 2016; Chen et al., 2017; Wang et al., 2017). In this paper, we attempt to introduce this idea to neural MT metrics.

## 3 Model

We wish to predict the score of a translation $t$ of length $l_t$ against a human reference $r$ of length $l_r$. For all models, we use fixed pre-trained contextualised word embeddings $\mathbf{e}_k$ to represent each word in the MT output and reference translation, in the form of matrices $\mathbf{W}_t$ and $\mathbf{W}_r$.

### 3.1 Unsupervised Model

We use cosine similarity to measure the pairwise similarity between $t$ and $r$ based on the maximum similarity score for each word embedding $\mathbf{e}_i \in t$ with respect to each word embedding $\mathbf{e}_j \in r$. We approximate recall of a word in $r$ with its maximum similarity with any word in $t$. The final predicted score, $y$, for a translation is the average recall of its reference:

$$\text{recall}_j = \max_{i=1}^{l_t} \text{cosine}(\mathbf{e}_i, \mathbf{e}_j) \quad (1)$$

$$y = \sum_{j=1}^{l_r} \frac{\text{recall}_j}{l_r} \quad (2)$$

### 3.2 Supervised Models

**Trained BiLSTM** We first encode the embeddings of the translation and reference with a bidirectional LSTM, and concatenate the max-pooled and average-pooled hidden states of the BiLSTM to generate $\mathbf{v_t}$ and $\mathbf{v_r}$, respectively:

$$\mathbf{v}_{s,max} = \max_{k=1}^{l_s} \mathbf{h}_{s,k}, \qquad \mathbf{v}_{s,avg} = \sum_{k=1}^{l_s} \frac{\mathbf{h}_{s,k}}{l_s} \quad (3)$$

$$\mathbf{v}_s = [\mathbf{v}_{s,max}; \mathbf{v}_{s,avg}] \quad (4)$$

To get the predicted score, we run a feedforward network over the concatenation of the sentence representations of $t$ and $r$, and their element-wise product and difference (a useful heuristic first proposed by Mou et al. (2016)). We train the model by minimizing mean squared error with respect to human scores.

$$\mathbf{m} = [\mathbf{v}_t; \mathbf{v}_r; \mathbf{v}_t \odot \mathbf{v}_r; \mathbf{v}_t - \mathbf{v}_r] \quad (5)$$

$$y = \mathbf{w}^\intercal \text{ReLU}(\mathbf{W}^\intercal \mathbf{m} + b) + b' \quad (6)$$

This is similar to RUSE, except that we learn the sentence representation instead of using pretrained sentence embeddings.

**Trained BiLSTM + attention** To obtain a sentence representation of the translation which is conditioned on the reference, we compute the attention-weighted representation of each word in the translation. The attention weights are obtained by running a softmax over the dot product similarity between the hidden state of the translation and reference BiLSTM. Similarly, we compute the relevant representation of the reference:

$$a_{i,j} = \mathbf{h}_{\mathbf{r}i}^\intercal \mathbf{h}_{\mathbf{t}j} \quad (7)$$

$$\tilde{\mathbf{h}_r} = \sum_{j=1}^{l_t} \frac{\exp(a_{i,j})}{\Sigma_i \exp(a_{i,j})} \cdot \mathbf{h}_t \quad (8)$$

$$\tilde{\mathbf{h}_t} = \sum_{i=1}^{l_r} \frac{\exp(a_{i,j})}{\Sigma_j \exp(a_{i,j})} \cdot \mathbf{h}_r \quad (9)$$

We then use $\tilde{\mathbf{h}}_t$ and $\tilde{\mathbf{h}}_r$ as our sentence representations in Eq. (3)–(6) to compute the final scores.

**Enhanced Sequential Inference Model (ESIM):** We also directly adapt ESIM (Chen et al., 2017), a high-performing model on the Natural Language Inference task, to the MT evaluation setting. We treat the human reference translation and the MT output as the premise and hypothesis, respectively.

The ESIM model first encodes $r$ and $t$ with a BiLSTM, then computes the attention-weighted representations of each with respect to the other (Eq. (7)–(9)). This model next "enhances" the representations of the translation (and reference) by capturing the interactions between $\mathbf{h}_t$ and $\tilde{\mathbf{h}}_t$ (and $\mathbf{h}_r$ and $\tilde{\mathbf{h}}_r$):

$$\mathbf{m}_r = [\mathbf{h}_r; \tilde{\mathbf{h}}_r; \mathbf{h}_r \odot \tilde{\mathbf{h}}_r; \mathbf{h}_r - \tilde{\mathbf{h}}_r] \quad (10)$$

$$\mathbf{m}_t = [\mathbf{h}_t; \tilde{\mathbf{h}}_t; \mathbf{h}_t \odot \tilde{\mathbf{h}}_t; \mathbf{h}_t - \tilde{\mathbf{h}}_t] \quad (11)$$

We use a feedforward projection layer to project these representations back to the model dimension, and then run a BiLSTM over each representation to compose local sequential information. The final representation of each pair of reference and translation sentences is the concatenation of the average-pooled and max-pooled hidden states of this BiLSTM. To compute the final predicted score, we apply a feedforward regressor over the concatenation of the two sentence representations.

$$\mathbf{p} = [\mathbf{v}_{r,avg}; \mathbf{v}_{r,max}; \mathbf{v}_{t,avg}; \mathbf{v}_{t,max}] \quad (12)$$

$$y = \mathbf{w}^\intercal \text{ReLU}(\mathbf{W}^\intercal \mathbf{p} + b) + b' \quad (13)$$

For all models, the predicted score of an MT system is the average predicted score of all its translations in the testset.

## 4 Experimental Setup

We use human evaluation data from the Conference on Machine Translation (WMT) to train and evaluate our models (Bojar et al., 2016, 2017a), which is based on the Direct Assessment ("DA") method (Graham et al., 2015, 2017). Here, system translations are evaluated by humans in comparison to a human reference translation, using a continuous scale (Graham et al., 2015, 2017). Each annotator assesses a set of 100 items, of which 30 items are for quality control, which is used to filter out annotators who are unskilled or careless. Individual worker scores are first standardised, and then the final score of an MT system is computed as the average score across all translations in the test set.

Manual MT evaluation is subjective and difficult, and it is not possible even for a diligent human to be entirely consistent on a continuous scale. Thus, any human annotations are noisy by nature. To obtain an accurate score for individual translations, the average score is calculated from scores of at least 15 "good" annotators. This data is then used to evaluate automatic metrics at the sentence level (Graham et al., 2015).

We train on the human evaluation data of news domain of WMT 2016, which is entirely crowdsourced. The sentence-level-metric evaluation data consists of accurate scores for 560 translations each for 6 to-English language pairs and English-to-Russian (we call this the "TrainS" dataset). The dataset also includes mostly singly-annotated[2] DA scores for around 125 thousand translations from six source languages into English, and 12.5 thousand translations from English-to-Russian ("TrainL" dataset), that were collected to obtain human scores for MT systems.

For the validation set, we use the sentence-level DA judgements collected for the WMT 2015 data (Bojar et al., 2015): 500 translation-reference pairs each of four to-English language pairs, and English-to-Russian.

For more details on implementation and training of our models, see Appendix A.

We test our metrics on all language pairs from the WMT 2017(Bojar et al., 2017b) news task in both the sentence and system level setting, and evaluate using Pearson's correlation between our metrics' predictions and the Human DA scores.

For the sentence level evaluation, insufficient DA annotations were collected for five from-English language pairs, and these were converted to preference judgements. If two MT system translations of a source sentence were evaluated by at least two reliable annotators, and the average score for System A is reasonably greater than the average score of System B, then this is interpreted as a Relative Ranking (DARR) judgement where Sys A is better than Sys B. The metrics are then evaluated using (a modified version of) Kendall's Tau correlation over these preference judgements.

We also evaluate on out-of-domain, system

[2]about 15% of the translations have a repeat annotation collected as part of quality-control

level data for five from-English language pairs from the WMT 2016 IT task.

## 5 Results

Tab. 1 compares the performance of our proposed metrics against existing metrics on the WMT 17 to-English news dataset. MEANT_2.0 (Lo, 2017) is the best untrained metric — it uses pre-trained word2vec embeddings (Mikolov et al., 2013)—, and RUSE (Shimanaka et al., 2018) is the best trained metric. We also include SENT-BLEU and CHRF baselines.

Our simple average recall metric ("BERTR") has a higher correlation than all existing metrics, and is highly competitive with RUSE. When trained on the sentence-level data (as with RUSE), the BiLSTM baseline does not perform well, however adding attention makes it competitive with RUSE. The ESIM model — which has many more parameters — underperforms compared to the BiLSTM model with attention.

However, the performance of all models improves substantially when these metrics are trained on the larger, singly-annotated training data (denoted "TrainL"), i.e., using data from only those annotators who passed quality control. Clearly the additional input instances make up for the increased noise level in the prediction variable. The simple BiLSTM model performs as well as RUSE, and both the models with attention substantially outperform this benchmark.

In this setting, we look at how the performance of ESIM improves as we increase the number of training instances (Fig. 1). We find that on the same number of training instances (3360), the model performs better on cleaner data compared to singly-annotated data ($r = 0.57$ vs $0.64$). However, when we have a choice between collecting multiple annotations for the same instances vs collecting annotations for additional instances, the second strategy leads to more gains.

We now evaluate the unsupervised BERTR model and the ESIM model (trained on the large dataset) in the other settings. In the sentence level tasks out-of-English (Tab. 4), the BERTR model (based on BERT-Chinese) significantly outperforms all metrics in the English-to-Chinese testset. For other language pairs, BERTR (based on multilingual BERT) is highly competitive with other metrics. ESIM performs well in the language pairs that are evaluated using Pearson's correlation. But the results are mixed when evaluated based on preference judgements. This could be an effect of our training method – using squared error as part of regression loss – being better suited to Pearson's $r$ — and might be resolved through a different loss, such as hinge loss over pairwise preferences which would better reflect Kendall's Tau (Stanojevic and Sima'an, 2014). Furthermore, ESIM is trained only on to-English and to-Russian data. It is likely that including more language pairs in the training data will increase correlation.

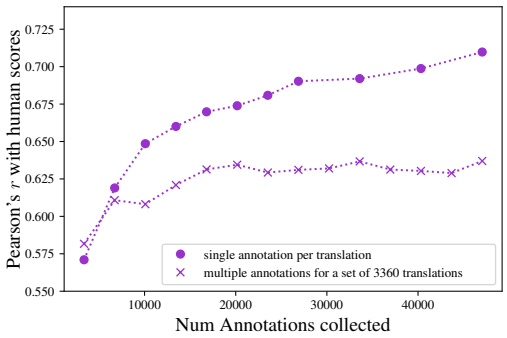


Figure 1: Average Pearson's $r$ for ESIM over the WMT 2017 to-English sentence-level dataset vs. the total number of annotations in the training set. We contrast two styles of collecting data: (1) the circles are trained on a single annotation per instance; and (2) the crosses are trained on the mean of N annotations per instance, as N goes from 1 to 14. The first strategy is more data-efficient.

On the system level evaluation of the news domain, both metrics are competitive with all other metrics in all language pairs both to- and out-of-English (see Tab. 3 and Tab. 4 in Appendix B).

In the IT domain, we have mixed results (Tab. 5 in the Appendix). ESIM significantly outperforms all other metrics in English–Spanish, is competitive in two other language pairs, and is outperformed by other metrics in the remaining two language pairs.

### 5.1 Qualitative Analysis

We manually inspect translations in the validation set. Tab. 6 in Appendix C shows examples of good translations, where our proposed metrics correctly recognise synonyms and valid word re-orderings, unlike SENT-BLEU. However, none of the metrics recognise a different way of expressing the same meaning. From Tab. 7, we see that SENT-BLEU gives high scores to translations with high partial overlap with the reference, but ESIM cor-

| | | cs–en | de–en | fi–en | lv–en | ru–en | tr–en | zh–en | **AVE.** |
|---|---|---|---|---|---|---|---|---|---|
| Baselines | BLEU | 0.435 | 0.432 | 0.571 | 0.393 | 0.484 | 0.538 | 0.512 | 0.481 |
| | CHRF | 0.514 | 0.531 | 0.671 | 0.525 | 0.599 | 0.607 | 0.591 | 0.577 |
| | MEANT_2.0 | 0.578 | 0.565 | 0.687 | 0.586 | 0.607 | 0.596 | 0.639 | 0.608 |
| | RUSE | 0.614 | 0.637 | 0.756 | 0.705 | 0.680 | 0.704 | 0.677 | 0.682 |
| P | BERTR | 0.655 | 0.650 | 0.777 | 0.671 | 0.680 | 0.702 | 0.687 | 0.689 |
| TrainS | BiLSTM | 0.517 | 0.556 | 0.735 | 0.672 | 0.606 | 0.619 | 0.565 | 0.610 |
| | BiLSTM + attention | 0.611 | 0.603 | 0.763 | 0.740 | 0.655 | 0.695 | 0.694 | 0.680 |
| | ESIM | 0.534 | 0.546 | 0.757 | 0.704 | 0.621 | 0.632 | 0.629 | 0.632 |
| TrainL | BiLSTM | 0.628 | 0.621 | 0.774 | 0.732 | 0.689 | 0.682 | 0.655 | 0.682 |
| | BiLSTM + attention | **0.704** | **0.710** | **0.818** | **0.777** | **0.744** | 0.753 | **0.737** | 0.749 |
| | ESIM | **0.692** | **0.706** | **0.829** | **0.764** | **0.726** | **0.776** | **0.732** | 0.746 |

Table 1: Pearson's $r$ on the WMT 2017 sentence-level evaluation data. P: Unsupervised metric that relies on pretrained embeddings; TrainS: trained on accurate 3360 instances; TrainL: trained on noisy 125k instances. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold (William's test; Graham and Baldwin, 2014)

| | | en–cs $\tau$ | en–de $\tau$ | en–fi $\tau$ | en–lv $\tau$ | en–ru $\rho$ | en–tr $\tau$ | en–zh $\rho$ |
|---|---|---|---|---|---|---|---|---|
| Baselines | SENT-BLEU | 0.274 | 0.269 | 0.446 | 0.259 | 0.468 | **0.377** | 0.642 |
| | CHRF | **0.376** | **0.336** | **0.503** | **0.420** | 0.605 | **0.466** | 0.608 |
| | BEER | **0.398** | **0.336** | **0.557** | **0.420** | 0.569 | **0.490** | 0.622 |
| | MEANT_2.0-NOSRL | **0.395** | **0.324** | **0.565** | **0.425** | 0.636 | **0.482** | 0.705 |
| | MEANT_2.0 | – | – | – | – | – | – | 0.727 |
| T P | BERTR | **0.390** | **0.365** | **0.564** | **0.417** | 0.630 | **0.457** | **0.803** |
| T | ESIM | 0.338 | **0.362** | **0.523** | 0.350 | **0.700** | **0.506** | 0.699 |

Table 2: Pearson's $r$ and Kendall's $\tau$ on the WMT 2017 from-English system-level evaluation data. The first section represents existing metrics, both trained and untrained. We then present results of our unsupervised metric, followed by our supervised metric trained in the TrainL setting: noisy 125k instances. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold (William's test (Graham and Baldwin, 2014) for Pearson's $r$ and Bootstrap (Efron and Tibshirani, 1993) for Kendall's $\tau$.)

rectly recognises then as low quality translations. However, in some cases, ESIM can be too permissive of bad translations which contain closely related words. There are also examples where a small difference in words completely changes the meaning of the sentence, but all the metrics score these translations highly.

## 6 Conclusion

We show that contextual embeddings are very useful for evaluation, even in simple untrained models, as well as in deeper attention based methods. When trained on a larger, much noisier range of instances, we demonstrate a substantial improvement over the state of the art.

In future work, we plan to extend these models by using cross-lingual embeddings, and combine information from translation–source interactions as well as translation–reference interactions. There are also direct applications to Quality Estimation, by using the source instead of the reference.

## Acknowledgements

We thank the anonymous reviewers for their feedback and suggestions to improve the paper. This work was supported in part by the Australian Research Council. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne.

## References


Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017a. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017b. Results of the wmt17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 489–513, Copenhagen, Denmark.

Julio Castillo and Paula Estrella. 2012. Semantic textual similarity for MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 52–58.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019)*, Minneapolis, USA.

Bradley Efron and Robert Tibshirani. 1993. *An introduction to the bootstrap*, volume 57. CRC press.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic features for automatic evaluation of heterogenous MT systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *EMNLP*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, pages 1183–1191, Denver, USA.

Rohit Gupta, Constantin Orasan, and Josef van Genabith. 2015. 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, Lisbon, Portugal.

Francisco Guzmán, Shafiq R Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 687–698.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. TESLA: Translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics (MATR)*, pages 354–359.

Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 589–597, Copenhagen, Denmark.

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, Volume 2: Shared Task Papers*, pages 598–603, Copenhagen, Denmark.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 130–136.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Chris Manning. 2009. Robust machine translation evaluation with entailment features. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 297–305.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, USA.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, and Phil Blunsom. 2016. Reasoning about entailment with neural attention. In *International Conference on Learning Representations (ICLR)*.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 764–771, Belgium, Brussels.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the Association for Machine Transaltion in the Americas*, pages 223–231.

Xingyi Song and Trevor Cohn. 2011. Regression and ranking based optimisation for sentence level machine translation evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 123–129.

Milos Stanojevic and Khalil Sima'an. 2014. BEER: BEtter evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419, Baltimore, USA.

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, Berlin, Germany.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4144–4150.

## A Implementation details

We implement our models using AllenNLP in PyTorch. We experimented with both ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) embeddings, and found that BERT consistently performs as well as, or better than ELMo, thus we report results using only BERT embeddings in this paper.

For BERTR, we use the top layer embeddings of the wordpieces of the MT and Reference translations. We use `bert_base_uncased` for all to-English language pairs, `bert_base_chinese` models for English-to-Chinese and `bert_base_multilingual_cased` for the remaining to-English language pairs.

For the trained metrics, we learn a weighted average of all layers of BERT embeddings. On the to-English testsets, we use `bert_base_uncased` embeddings and train on the WMT16 to-English data. On all other testsets, we use the `bert_base_multilingual_cased` embeddings and train on the WMT 2016 English-to-Russian, as well as all to-English data.

Following the recommendations of the original ESIM paper, we fix the dimension of the BiLSTM hidden state to 300 and set the Dropout rate to 0.5. We use the Adam optimizer with an initial learning rate of 0.0004 and batch size of 32, and use early stopping on the validation dataset.

Training the ESIM model on the full dataset takes around two hours on a single V100 GPU, and all models take less than two minutes to evaluate a standard WMT dataset of 3000 translations.

## B System-level results for WMT 17 news and WMT 2016 IT domain

| | | cs–en | de–en | fi–en | lv–en | ru–en | tr–en | zh–en |
|---|---|---|---|---|---|---|---|---|
| | num systems | 4 | 11 | 6 | 9 | 9 | 10 | 16 |
| Baselines | BLEU | 0.971 | 0.923 | 0.903 | **0.979** | 0.912 | 0.976 | 0.864 |
| | CHRF | 0.939 | **0.968** | **0.938** | 0.968 | 0.952 | 0.944 | 0.859 |
| | CHARACTER | 0.972 | **0.974** | **0.946** | 0.932 | **0.958** | 0.949 | 0.799 |
| | BEER | 0.972 | **0.960** | **0.955** | **0.978** | 0.936 | 0.972 | 0.902 |
| | RUSE | **0.990** | **0.968** | **0.977** | 0.962 | **0.953** | **0.991** | **0.974** |
| TP | BERTR | **0.996** | **0.971** | **0.948** | **0.980** | **0.950** | **0.994** | **0.970** |
| T | ESIM | 0.983 | **0.949** | **0.985** | **0.974** | **0.921** | **0.986** | 0.901 |

Table 3: Pearson's $r$ on the WMT 2017 to-English system-level evaluation data. The first section represents existing metrics, both trained and untrained. We then present results of our unsupervised metric, followed by our supervised metric trained in the TrainL setting: noisy 130k instances. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

| | | en–cs | en–de | en–fi | en–lv | en–ru | en–tr | en–zh |
|---|---|---|---|---|---|---|---|---|
| | num systems | 14 | 16 | 12 | 17 | 9 | 8 | 11 |
| Baselines | BLEU | 0.956 | 0.804 | 0.920 | 0.866 | 0.898 | 0.924 | 0.981 |
| | BEER | **0.970** | 0.842 | 0.976 | **0.930** | 0.944 | 0.980 | 0.914 |
| | CHARACTER | **0.981** | **0.938** | **0.972** | 0.897 | **0.939** | 0.975 | 0.933 |
| | CHRF | **0.976** | 0.863 | **0.981** | **0.955** | 0.950 | **0.991** | 0.976 |
| TP | BERTR | **0.982** | 0.877 | **0.979** | 0.949 | **0.971** | **0.996** | **0.992** |
| T | ESIM | **0.974** | 0.861 | **0.971** | **0.954** | **0.968** | 0.978 | 0.970 |

Table 4: Pearson's $r$ on the WMT 2017 from-English system-level evaluation data. The first section represents existing metrics, both trained and untrained. We then present results of our unsupervised metric, followed by our supervised metric trained in the TrainL setting: noisy 130k instances. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

| | | en–cs | en–de | en–es | en–nl | en–pt |
|---|---|---|---|---|---|---|
| | num systems | 5 | 10 | 4 | 4 | 4 |
| Baselines | BLEU | 0.750 | 0.621 | 0.976 | 0.596 | **0.997** |
| | CHRF | **0.845** | 0.588 | 0.915 | **0.951** | 0.967 |
| | BEER | **0.744** | 0.621 | 0.931 | **0.983** | 0.989 |
| | CHARACTER | **0.901** | **0.930** | 0.963 | **0.927** | 0.976 |
| TP | BERTR | **0.974** | 0.780 | **0.925** | 0.896 | 0.980 |
| T | ESIM | **0.964** | 0.780 | **0.991** | 0.798 | 0.996 |

Table 5: Pearson's $r$ on the WMT 2016 IT domain system-level evaluation data. The first section represents existing metrics, both trained and untrained. We then present results of our pretrained metric, followed by our supervised metric trained in the TrainL setting: noisy 130k instances. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

## C Qualitative analysis

| | Translations with HIGH Human scores | ESIM | BERTR | SENT-BLEU |
|---|---|---|---|---|
| ref: | The negotiations have been scheduled to take place next Saturday, the Russian Minister of Energy, Alexander Nowak, said on Monday. | HIGH | HIGH | LOW |
| sys: | The negotiations are scheduled for coming Saturday, said the Russian energy minister Alexander Nowak on Monday. | | | |
| ref: | Lesotho military says no coup planned; PM stays in South Africa | | | |
| sys: | Lesotho-military member says that no coup is planned; Prime Minister remains in South Africa | | | |
| ref: | In September 2011, Abbott's condition worsened again, and his consultant took his CT scans and X-rays to a panel of experts. | | | |
| sys: | In September 2011 Abbotts state worsened again and his family doctor brought his CT-Scans and X-rays to an expert group. | | | |
| ref: | The boardroom is now contemplating the possibility of working together. | HIGH | LOW | LOW |
| sys: | Now the boards are thinking about a possible cooperation. | | | |
| ref: | He ended up spending a month off work. | LOW | LOW | LOW |
| sys: | In the end, he could not go to work for a month. | | | |

Table 6: Examples of good translations in the WMT 2015 sentence level DA dataset and whether ESIM, BERTR and SENT-BLEU correctly give them high scores

| | Translations with LOW Human scores | ESIM | BERTR | SENT-BLEU |
|---|---|---|---|---|
| ref: | For the benefit of the school, Richter nurtured a good relationship with the then Mayor, Ludwig Gtz (CSU). | LOW | LOW | HIGH |
| sys: | For the good of the school of judges as rector of a good relationship with the former mayor Ludwig Gtz (CSU) | | | |
| ref: | The military plays an important role in Pakistan and has taken power by force several times in the past. | | | |
| sys: | The military plays an important role in Pakistan and has already more frequently geputscht. | | | |
| ref: | Behind much of the pro-democracy campaign in Hong Kong is the Occupy Central With Love and Peace movement, whose organizers have threatened to shut down the financial district if Beijing does not grant authentic universal suffrage. | LOW | HIGH | HIGH |
| sys: | Behind the pro-democracy campaign in Hong Kong is the movement Occupy Central With Love and Peace, whose organizers have threatened the acupuncture, off, if Beijing allows no real universal suffrage. | | | |
| ref: | Foreign goods trade had slowed, too. | HIGH | LOW | LOW |
| sys: | Foreign trade also slowed the economy. | | | |
| ref: | Some shrapnel pieces are still in my knee. | HIGH | HIGH | HIGH |
| sys: | Some garnet fragments are still in my knee. | | | |
| ref: | Stewart hit the wall for the second time after his right front tire blew out on lap 172, ending his night. | | | |
| sys: | Stewart raced for the second time against the wall after his right front tire on lap 172 and ended his evening. | | | |

Table 7: Examples of bad quality translations in the WMT 2015 sentence level DA dataset and whether ESIM, BERTR and SENT-BLEU correctly give them low scores