Beyond BLEU:
Training Neural Machine Translation with Semantic Similarity

John Wieting[1], Taylor Berg-Kirkpatrick[2], Kevin Gimpel[3], and Graham Neubig[1]
[1]Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[2]University of California San Diego, San Diego, CA, 92093, USA
[3]Toyota Technological Institute at Chicago, Chicago, IL, 60637, USA
{jwieting,gneubig}@cs.cmu.edu, tberg@eng.ucsd.edu, kgimpel@ttic.edu

Abstract

While most neural machine translation (NMT) systems are still trained using maximum likelihood estimation, recent work has demonstrated that optimizing systems to directly improve evaluation metrics such as BLEU can substantially improve final translation accuracy. However, training with BLEU has some limitations: it doesn't assign partial credit, it has a limited range of output values, and it can penalize semantically correct hypotheses if they differ lexically from the reference. In this paper, we introduce an alternative reward function for optimizing NMT systems that is based on recent work in semantic similarity. We evaluate on four disparate languages translated to English, and find that training with our proposed metric results in better translations as evaluated by BLEU, semantic similarity, and human evaluation, and also that the optimization procedure converges faster. Analysis suggests that this is because the proposed metric is more conducive to optimization, assigning partial credit and providing more diversity in scores than BLEU.[1]

1 Introduction

In neural machine translation (NMT) and other natural language generation tasks, it is common practice to improve likelihood-trained models by further tuning their parameters to explicitly maximize an automatic metric of system accuracy – for example, BLEU (Papineni et al., 2002) or METEOR (Denkowski and Lavie, 2014). Directly optimizing accuracy metrics involves backpropagating through discrete decoding decisions, and thus is typically accomplished with structured prediction techniques like reinforcement learning (Ranzato et al., 2016), minimum risk training (Shen

et al., 2015), and other specialized methods (Wiseman and Rush, 2016). Generally, these methods work by repeatedly generating a translation under the current parameters (via decoding, sampling, or loss-augmented decoding), comparing the generated translation to the reference, receiving some reward based on their similarity, and finally updating model parameters to increase future rewards.

In the vast majority of work, discriminative training has focused on optimizing BLEU (or its sentence-factored approximation). This is not surprising given that BLEU is the standard metric for system comparison at test time. However, BLEU is not without problems when used as a training criterion. Specifically, since BLEU is based on n-gram precision, it aggressively penalizes lexical differences even when candidates might be synonymous with or similar to the reference: if an n-gram does not exactly match a sub-sequence of the reference, it receives no credit. While the pessimistic nature of BLEU differs from human judgments and is therefore problematic, it may, in practice, pose a more substantial problem for a different reason: BLEU is difficult to *optimize* because it does not assign partial credit. As a result, learning cannot hill-climb through intermediate hypotheses with high synonymy or semantic similarity, but low n-gram overlap. Furthermore, where BLEU does assign credit, the objective is often flat: a wide variety of candidate translations can have the same degree of overlap with the reference and therefore receive the same score. This, again, makes optimization difficult because gradients in this region give poor guidance.

In this paper we propose SIMILE, a simple alternative to matching-based metrics like BLEU for use in discriminative NMT training. As a new reward, we introduce a measure of *semantic similarity between the generated hypotheses and the reference translations* evaluated by an embed-

---

ding model trained on a large external corpus of paraphrase data. Using an embedding model to evaluate similarity allows the range of possible scores to be continuous and, as a result, introduces fine-grained distinctions between similar translations. This allows for partial credit and reduces the penalties on semantically correct but lexically different translations. Moreover, since the output of SIMILE is continuous, it provides more informative gradients during the optimization process by distinguishing between candidates that would be similarly scored under matching-based metrics like BLEU. Lastly, we show in our analysis that SIMILE has an additional benefit over BLEU by translating words with heavier semantic content more accurately.

To define an exact metric, we reference the burgeoning field of research aimed at measuring semantic textual similarity (STS) between two sentences (Le and Mikolov, 2014; Pham et al., 2015; Wieting et al., 2016; Hill et al., 2016; Conneau et al., 2017; Pagliardini et al., 2017). Specifically, we start with the method of Wieting and Gimpel (2018), which learns paraphrastic sentence representations using a contrastive loss and a parallel corpus induced by backtranslating bitext. Wieting and Gimpel showed that simple models that average word or character trigram embeddings can be highly effective for semantic similarity. The strong performance, domain robustness, and computationally efficiency of these models make them highly attractive. For the purpose of discriminative NMT training, we augment these basic models with two modifications: we add a length penalty to avoid short translations, and compose the embeddings of subword units, rather than words or character trigrams, in order to compute similarity. We find that using subword units also yields better performance on the STS evaluations and is more efficient than character trigrams.

We conduct experiments with our new metric on the 2018 WMT (Bojar et al., 2018) test sets, translating four languages, Czech, German, Russian, and Turkish, into English. Results demonstrate that optimizing SIMILE during training results in not only improvements in the same metric during test, but also in consistent improvements in BLEU. Further, we conduct a human study to evaluate system outputs and find significant improvements in human-judged translation quality for all but one language. Finally, we provide an analysis

of our results in order to give insight into the observed gains in performance. Tuning for metrics other than BLEU has not (to our knowledge) been extensively examined for NMT, and we hope this paper provides a first step towards broader consideration of training metrics for NMT.

2 SIMILE Reward Function

Since our goal is to develop a continuous metric of sentence similarity, we borrow from a line of work focused on domain agnostic semantic similarity metrics. We motivate our choice for applying this line of work to training translation models in Section 2.1. Then in Section 2.2, we describe how we train our similarity metric (SIM), how we compute our length penalty, and how we tie these two terms together to form SIMILE.

2.1 SIMILE

Our SIMILE metric is based on the sentence similarity metric of Wieting and Gimpel (2018), which we choose as a starting point because it has state-of-the-art unsupervised performance on a host of domains for semantic textual similarity.[2] Both being unsupervised and fairly domain agnostic imply that it generalizes well to unseen examples in contrast to supervised methods which are often imbued with the bias of their training data.

Model. Our sentence encoder g averages 300 dimensional subword unit[3] embeddings to create a sentence representation. The similarity of two sentences, SIM, is obtained by encoding both with g and then calculating their cosine similarity.

Training. We follow Wieting and Gimpel (2018) in learning the parameters of the encoder g. The training data is a set S of paraphrase pairs[4] $\langle s, s' \rangle$ and we use a margin-based loss:

$$\ell(s, s') = \max(0, \delta - \cos(g(s), g(s')) + \cos(g(s), g(t)))$$

Model	2012	2013	2014	2015	2016
SIM	**69.3**	**64.1**	**77.2**	**80.3**	**78.6**
SIMILE	70.1	59.8	74.7	79.4	77.8
Wieting and Gimpel (2018)	67.8	62.8	76.9	79.8	76.8
BLEU	39.2	29.5	42.8	49.8	47.4
METEOR	53.4	47.6	63.7	68.8	61.8
STS 1st Place	64.8	62.0	74.3	79.0	77.7
STS 2nd Place	63.4	59.1	74.2	78.0	75.7
STS 3rd Place	64.1	58.3	74.3	77.8	75.7

Table 1: Comparison of the semantic similarity model used in this paper (SIM) with a number of strong baselines including the model of (Wieting and Gimpel, 2018) and the top 3 performing STS systems for each year.

Model	newstest2015	newstest2016
SIM	58.2	53.1
SIMILE	58.4	53.2
BLEU	13.0	9.7
METEOR	**58.9**	**57.2**

Table 2: Comparison of models on machine translation quality evaluation datasets. Scores are in Spearman's ρ.

where δ is the margin, and t is a *negative example* taken from a mini-batch during optimization. The intuition is that we want the two texts to be more similar to each other than to their negative examples. To select t we choose the most similar sentence in a collection of mini-batches called a *mega-batch*.

Finally, we note that SIM is robust to domain, as shown by its strong performance on the STS tasks which cover a broad range of domains. Also, although SIM was trained primarily on subtitles, we use news data to train and evaluate our NMT models, showing improved performance over a baseline using BLEU.

Length Penalty. Our initial experiments showed that when using just the similarity metric, SIM, there was nothing preventing the model from learning to generate long sentences, often at the expense of repeating words. This is the opposite case from BLEU, where the n-gram precision is not penalized for generating too few words. Therefore, in BLEU, a brevity penalty (BP) was introduced to penalize sentences when they are shorter than the reference. The penalty is:

$$\text{BP}(r, h) = e^{1 - \frac{|r|}{|h|}}$$

where r is the reference and h is the generated hypothesis, with $|r|$ and $|h|$ their respective lengths. We experimented with modifying this penalty to only penalize generated sentences that are longer than the target (so we switch r and h in the equa-

tion). However, we found that this favored short sentences. We instead penalize a generated sentence if its length differs at all from that of the target. Therefore, our length penalty is:

$$\text{LP}(r, h) = e^{1 - \frac{\max(|r|, |h|)}{\min(|r|, |h|)}}$$

SIMILE. Our final metric, which we refer to as SIMILE, is defined as follows:

$$\text{SIMILE} = \text{LP}(r, h)^{\alpha} \text{SIM}(r, h)$$

In initial experiments we found that performance could be improved slightly by lessening the influence of LP, so we tune α over the set $\{0.25, 0.5\}$. Overall, our results were robust to the choice of α, but there was some benefit from tuning over these two values.

2.2 Motivation

There is a vast literature on metrics for *evaluating* machine translation outputs automatically (For instancem the WMT metrics task papers like Bojar et al. (2017)). In this paper we demonstrate that *training* towards metrics other than BLEU has significant practical advantages in the context of NMT. While this could be done with any number of metrics, in this paper we experiment with a single semantic similarity metric, and due to resource constraints leave a more extensive empirical comparison of other evaluation metrics to future work. That said, we designed SIMILE as a semantic similarity model with high accuracy, domain robustness, and computational efficiency to be used in minimum risk training for machine translation.[5]

While semantic similarity is not an exact replacement for measuring machine translation quality, we argue that it serves as a decent proxy at least as far as minimum risk training is concerned. To test this, we compare the similarity metric term in SIMILE (SIM) to BLEU and METEOR on two machine quality datasets[6] and report their correlation with human judgments in Table 2. Machine translation quality measures account for more than semantics as they also capture other factors like fluency. A manual error analysis and the fact that the machine translation correlations in Table 2 are

---

[5] SIMILE, including time to split the sentence is about 20 times faster than METEOR when code is executed on GPU (NVIDIA GeForce GTX 1080).

[6] We used the segment level data from newstest2015 and newstest2016 available at http://statmt.org/wmt18/metrics-task.html. The former contains 7 language pairs and the latter 5.

close, but the semantic similarity correlations[7] in Table 1 are not, suggest that the difference between METEOR and SIM largely lies in fluency. However, not capturing fluency is something that can be ameliorated by adding a down-weighted maximum-likelihood (MLE) loss to the minimum risk loss. This was done by Edunov et al. (2018) and we use this in our experiments as well.

3 Machine Translation Preliminaries

Architecture. Our model and optimization procedure are based on prior work on structured prediction training for neural machine translation (Edunov et al., 2018) and are implemented in Fairseq.[8] Our architecture follows the paradigm of an encoder-decoder with soft attention (Bahdanau et al., 2015) and we use the same architecture for each language pair in our experiments. We use gated convolutional encoders and decoders (Gehring et al., 2017). We use 4 layers for the encoder and 3 with the decoder, setting the hidden state size for all layers to 256, and the filter width of the kernels to 3. We use byte pair encoding (Sennrich et al., 2015), with a vocabulary size of 40,000 for the combined source and target vocabulary. The dimension of the BPE embeddings is also set to 256.

Objective Functions. Following (Edunov et al., 2018), we first train models with maximum-likelihood with label-smoothing ($\mathcal{L}_{\text{TokLS}}$) (Szegedy et al., 2016; Pereyra et al., 2017). We set the confidence penalty of label smoothing to be 0.1. Next, we fine-tune the model with a weighted average of minimum risk training ($\mathcal{L}_{\text{Risk}}$) (Shen et al., 2015) and ($\mathcal{L}_{\text{TokLS}}$), where the expected risk is defined as:

$$\mathcal{L}_{\text{Risk}} = \sum_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} \text{cost}(\mathbf{t}, \mathbf{u}) \frac{p(\mathbf{u}|\mathbf{x})}{\sum_{\mathbf{u}' \in \mathcal{U}(\mathbf{x})} p(\mathbf{u}'|\mathbf{x})}$$

where $\mathbf{u}$ is a candidate hypothesis, $\mathcal{U}(\mathbf{x})$ is a set of candidate hypotheses, and $\mathbf{t}$ is the reference.

---

[7]Evaluation is on the SemEval Semantic Textual Similarity (STS) datasets from 2012-2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016). In the SemEval STS competitions, teams create models that need to work well on domains both represented in the training data and hidden domains revealed at test time. Our model and those of Wieting and Gimpel (2018), in contrast to the best performing STS systems, do not use any manually-labeled training examples nor any other linguistic resources beyond the ParaNMT corpus (Wieting and Gimpel, 2018).

[8]Available at https://github.com/pytorch/fairseq.

Lang.	Train	Valid	Test
cs-en	218,384	6,004	2,983
de-en	284,286	7,147	2,998
ru-en	235,159	7,231	3,000
tr-en	207,678	7,008	3,000

Table 3: Number of sentence pairs in the training/validation/test sets for all four languages.

Therefore, our fine-tuning objective becomes:

$$\mathcal{L}_{\text{Weighted}} = \gamma \mathcal{L}_{\text{TokLS}} + (1 - \gamma) \mathcal{L}_{\text{Risk}}$$

We tune γ from the set $\{0.2, 0.3\}$ in our experiments. In minimum risk training, we aim to minimized the expected cost. In our case that is $1 - \text{BLEU}(t, h)$ or $1 - \text{SIMILE}(t, h)$ where t is the target and h is the generated hypothesis. As is commonly done, we use a smoothed version of BLEU by adding 1 to all n-gram counts except unigram counts. This is to prevent BLEU scores from being overly sparse (Lin and Och, 2004). We generate candidates for minimum risk training from n-best lists with 8 hypotheses without and do not include the reference in the candidates.

Optimization. We optimize our models using Nesterov's accelerated gradient method (Sutskever et al., 2013) using a learning rate of 0.25 and momentum of 0.99. Gradients are renormalized to norm 0.1 (Pascanu et al., 2012). We train the $\mathcal{L}_{\text{TokLS}}$ objective for 200 epochs and the combined objective, $\mathcal{L}_{\text{Weighted}}$, for 10. Model selection is done by selecting the model with the lowest validation loss on the validation set. Then, depending on the evaluation being considered, we select models with the highest performance on the validation set.

4 Experiments

4.1 Data

Training models with minimum risk is expensive, but we wanted to evaluate in a difficult, realistic setting using a diverse set of languages. Therefore, we experiment on four language pairs: Czech (cs-en), German (de-en), Russian (ru-en), and Turkish (tr-en) translating to English (EN). For training data for cs-en, de-en, and ru-en, we use News Commentary v13[9] provided by WMT (Bojar et al., 2018) for training the models. For training the Turkish system, we used the

---

[9]Available at http://data.statmt.org/wmt18/translation-task/training-parallel-nc-v13.tgz.

Model	de-en		cs-en		ru-en		tr-en	
	BLEU	SIM	BLEU	SIM	BLEU	SIM	BLEU	SIM
MLE	27.52	74.96	17.02	67.18	17.92	70.24	14.47	63.52
BLEU	27.95‡	86.93‡	17.29‡	81.92‡	17.92	84.63‡	15.00‡	80.30‡
SIMILE	**28.28**†‡	**87.32**†‡	**17.51**†‡	**82.12**†‡	18.23†‡	**85.12**†‡	15.28†‡	**81.04**†‡
Half	28.24†‡	87.11†‡	17.50†‡	82.11†‡	**18.24**†‡	85.10†‡	**15.34**†‡	80.64†‡

Table 4: Results on translating four languages to English for MLE, BLEU, SIMILE and Half. † denotes statistical significance ($p < 0.05$) over BLEU and ‡ denotes statistical significance over MLE. Statistical significance was computed using paired bootstrap resampling.

WMT 2018 parallel data which consisted of the SETIMES2[10] corpus. The validation and development sets for `de-en`, `cs-en`, and `ru-en` were the WMT 2016 and WMT 2017 validation sets. For `tr-en`, the validation set was the WMT 2016 validation set and WMT 2017 validation and test sets. Test sets for each language were the official WMT 2018 test sets.

4.2 Automatic Evaluation

We first use corpus-level BLEU and the corpus average SIM score to evaluate the outputs of the different experiments. It is important to note that in this case, SIM is not the same as SIMILE. SIM is only the semantic similarity component of SIMILE and therefore lacks the length penalization term. We used this metric to estimate the degree to which the semantic content of a translation and its reference overlap. When evaluating semantic similarity, we find that SIM outperforms SIMILE (by about a point across each year of STS tasks).

We compare systems trained with 4 objectives:

- MLE: Maximum likelihood with label smoothing
- BLEU: Minimum risk training with 1-BLEU as the cost
- SIMILE: Minimum risk training with 1-SIMILE as the cost
- Half: Minimum risk training with a new cost that is half BLEU and half SIMILE: $1 - \frac{1}{2}(\text{BLEU} + \text{SIM})$

The results are shown in Table 4. From the table, we see that using SIMILE performs the best when using BLEU and SIM as evaluation metrics for all four languages. It is interesting that using SIMILE in the cost leads to larger BLEU improvements than using BLEU alone, the reasons for which we examine further in the following sections. It is important to emphasize that increasing

Lang.	Avg. Score		
	MLE	BLEU	SIMILE
cs-en	0.98	0.90	**1.02**†
de-en	0.93	0.85	**1.00**†
ru-en	1.22	1.21	**1.31**†‡
tr-en	0.98*	**1.03***	0.78

Table 5: Average human ratings on 200 sentences from the test set for each of the respective languages. † denotes statistical significance ($p < 0.05$) over BLEU, except for the case of `cs-en`, where $p = 0.06$. ‡ denotes statistical significance over MLE, and * denotes statistical significance over SIMILE. Statistical significance was computed using paired bootstrap resampling.

BLEU was not the goal of our proposed method, human evaluations were our target, but this is a welcome surprise. Similarly, using BLEU as the cost function leads to large gains in SIM, though these gains are not as large as when using SIMILE in training.

4.3 Human Evaluation

We also perform human evaluation, comparing MLE training with minimum risk training using SIMILE and BLEU as costs. We selected 200 sentences along with their translation from the respective test sets of each language. The sentences were selected nearly randomly with the only constraints that they be between 3 and 25 tokens long and also that the outputs for SIMILE and BLEU were not identical. The translators then assigned a score from 0-5 based on how well the translation conveyed the information contained in the reference.[11]

From the table, we see that minimum risk training with SIMILE as the cost scores the highest across all language pairs except Turkish. Turkish is also the language with the lowest test BLEU (See Table 4). An examination of the human-annotated outputs shows that in Turkish (unlike the other languages) repetition was a significant

problem for the SIMILE system in contrast to MLE or BLEU. We hypothesize that one weakness of SIMILE may be that it needs to start with some minimum level of translation quality in order to be most effective. The biggest improvement over BLEU is on `de-en` and `ru-en`, which have the highest MLE BLEU scores in Table 4 which further lends credence to this hypothesis.

5 Quantitative Analysis

We next analyze our model using primarily the validation set of the `de-en` data. We chose this dataset for the analysis since it had the highest MLE BLEU scores of the languages studied.

5.1 Partial Credit

We analyzed the distribution of the cost function for both SIMILE and BLEU. Again, using a beam size of 8, we computed the cost for all generated translations and plotted their histogram in Figure 1.

The plots show that the distribution of scores for SIMILE and BLEU are quite different. Both distributions are not symmetrical Gaussian, however the distribution of BLEU scores is significantly more skewed with much higher costs. This tight clustering of costs provides less information during training.

Next, for all n-best lists, we computed all differences between scores of the hypotheses in the beam. Therefore, for a beam size of 8, this results in 28 different scores. We found that of the 86,268 scores, the difference between scores in an n-best list is ≥ 0 99.0% of the time for SIMILE, but 85.1% of the time for BLEU. The average difference is 4.3 for BLEU and 4.8 for SIMILE, showing that SIMILE makes finer grained distinctions among candidates.

5.2 Validation Loss

We next analyze the validation loss during training of the `de-en` model for both using SIMILE and BLEU as costs. We use the hyperparameters of the model with the highest BLEU on the validation set for model selection. Since the distributions of costs vary significantly between SIMILE and BLEU, with BLEU having much higher costs on average, we compute the validation loss with respect to both cost functions for each of the two models.

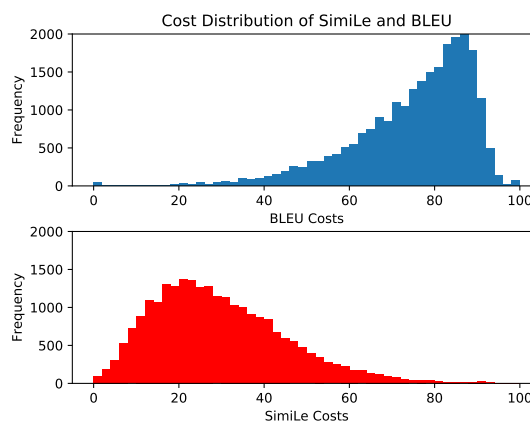In Figure 2, we plot the risk objective for each

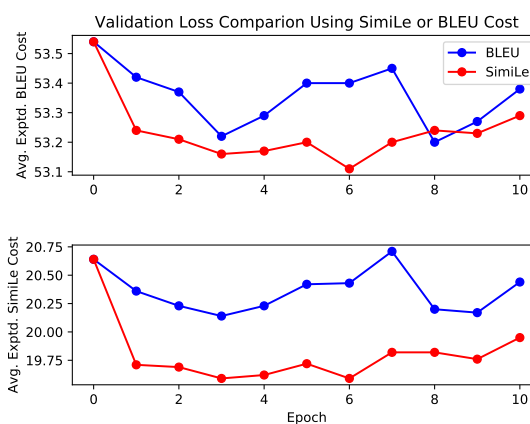Figure 1: Distribution of scores for SIMILE and BLEU.

Figure 2: Validation loss comparison for SIMILE and BLEU. The top plot shows the expected BLEU cost when training with BLEU and SIMILE. The bottom plot shows the expected SIMILE cost when training with BLEU and SIMILE.

of the 10 epochs during training. In the top plot, we see that the risk objective for both BLEU and SIMILE decreases much faster when using SIMILE to train than BLEU. The expected BLEU also reaches a significantly lower value on the validation set when training with SIMILE. The same trend occurs in the lower plot, this time measuring the expected SIMILE cost on the validation set.

From these plots, we see that optimizing with SIMILE results in much faster training. It also reaches a lower validation loss and from Table 4 we've already shown that the SIMILE and BLEU on the test set are higher for models trained with SIMILE. To hammer home the point at how much faster the models trained with SIMILE reach better performance, we evaluated just after 1 epoch of training and found that the model trained

with BLEU had SIM/BLEU scores of 86.71/27.63 while the model trained with SIMILE had scores of 87.14/28.10. A similar trend was observed in the other language pairs as well, where the validation curves show a much larger drop-off after a single epoch when training with SIMILE than with BLEU.

5.3 Effect of n-best List Size

As mentioned in Section 3, we used an n-best list size of 8 in our minimum risk training experiments. In this section, we train de-en translation models with various n-best list sizes and investigate the relationship between beam size and using SIMILE or BLEU as a cost. We hypothesize that since BLEU is not as fine-grained a metric as SIMILE, expanding the number of candidates would close the gap between BLEU and SIMILE as BLEU would have access to a more candidates with more diverse scores. The results of our experiment are shown in Figure 3 and show that models trained with SIMILE actually improve in BLEU and SIM more significantly as n-best list size increases. This is possibly due to small n-best sizes inherently upper-bounding performance regardless of training metric, and SIMILE being a better measure overall when the n-best is sufficiently large to learn.
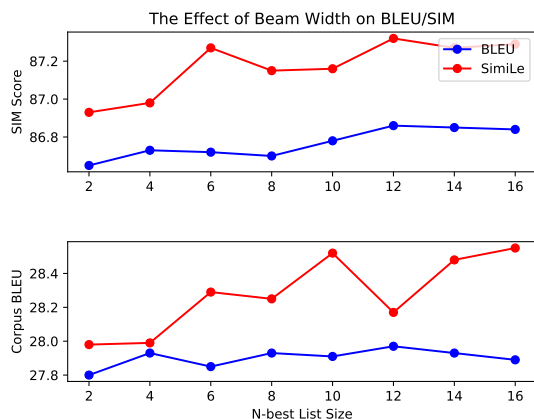

Figure 3: The relationship between n-best list size and performance as measured by Avg. SIM over the dataset or corpus-level BLEU when training using SIMILE or BLEU as a cost.

5.4 Lexical F1

We next attempt to elucidate exactly which parts of the translations are improving due to using SIMILE cost compared to using BLEU. We use

Lang./Bucket	cs-en Δ	de-en Δ	ru-en Δ	tr-en Δ	Avg.
1	0.1	0.8	0.2	0.1	0.30
2-5	1.2	0.6	*0.0*	0.2	0.50
6-10	0.4	0.7	1.4	*-0.3*	0.55
11-100	0.2	0.6	0.6	0.4	0.45
101-1000	*-0.3*	0.3	0.4	0.2	0.15
1001+	*-0.2*	0.5	0.4	*-0.0*	0.08
DET	0.1	*-0.1*	0.7	-0.5	0.03
PRON	0.6	*-0.3*	0.1	0.9	0.33
PREP	0.2	*-0.3*	0.5	0.5	0.24
CONJ	0.1	1.1	0.3	-0.5	0.27
PUNCT	*-0.4*	1.3	0.8	-0.4	0.34
NUM	0.6	2.2	1.8	1.3	**1.48**
SYM	0.3	3.6	4.4	1.7	**2.50**
INTJ	3.2	*-1.1*	3.2	-2.6	**0.66**
VERB	0.2	0.3	0.0	*0.0*	0.13
ADJ	0.2	0.7	0.3	-0.2	0.25
ADV	*-0.2*	0.1	0.8	0.7	0.34
NOUN	0.3	1.1	0.8	0.4	**0.63**
PRNOUN	0.5	1.2	0.6	0.4	**0.65**

Table 6: F1 score for various buckets of words. The values in the table are the difference between F1 for that specific language type and bucket between training using SIMILE and BLEU (positive values means SIMILE had a higher F1). The first part of the table shows F1 scores across bins defined by word frequency on the test set. So words appearing only 1 time are in the first row, between 2-5 times are in the second row, etc. The next part of the table buckets words by coarse POS tags.

compare-mt (Neubig et al., 2019)[12] to compute the F1 scores for target word types based on their frequency and their coarse part-of-speech-tag (as labeled by SpaCy[13]) and show the results in Table 6.

From the table, we see that training with SIMILE helps produce low frequency words more accurately, a fact that is consistent with the POS tag analysis in the second part of the table. Wieting and Gimpel (2017) noted that highly discriminative parts-of-speech, such as nouns, proper nouns, and numbers, made the most contribution to the sentence embeddings. Other works (Pham et al., 2015; Wieting et al., 2016) have also found that when training semantic embeddings using an averaging function, embeddings that bear the most information regarding the meaning have larger norms.

We also see that these same parts-of-speech (nouns, proper nouns, numbers) have the largest difference in F1 scores between SIMILE and BLEU. Other parts-of-speech like SYM and INTJ have high F1 scores as well, and words belonging to these classes are both relatively rare and highly discriminative regarding the semantics of the sen-

---

[12] Available at https://github.com/neulab/compare-mt.

[13] Available at https://github.com/explosion/spaCy.

Reference	System	Human Score	Translation
I will tell you my personal opinion of him	BLEU	2	I will have a personal opinion on it.
	SIMILE	4	I will tell my personal opinion about it.
	MLE	2	I will have a personal view of it.
In my case, it was very varied.	BLEU	0	I was very different from me.
	SIMILE	4	For me, it was very different.
	MLE	1	In me, it was very different.
We're making the city liveable.	BLEU	0	We make the City of Life Life.
	SIMILE	3	We make the city viable.
	MLE	0	We make the City of Life.
The head of the White House said that the conversation was ridiculous.	BLEU	0	The White House chairman, the White House chip called a ridiculous.
	SIMILE	4	The White House's head, he described the conversation as ridiculous.
	MLE	1	The White House chief, he called the White House, he called a ridiculous.
According to the former party leaders, so far the discussion has been predominated by expressions of opinion based on emotions, without concrete arguments.	BLEU	3	According to former party leaders, the debate has so far had to be "elevated to an expression of opinion without concrete arguments."
	SIMILE	5	In the view of former party leaders, the debate has been based on emotions without specific arguments."
	MLE	4	In the view of former party leaders, in the debate, has been based on emotions without specific arguments."
We are talking about the 21st century: servants.	BLEU	4	We are talking about the 21st century: servants.
	SIMILE	1	In the 21st century, the 21st century is servants.
	MLE	0	In the 21st century, the 21st century is servants.
Prof. Dr. Caglar continued:	BLEU	3	They also reminded them.
	SIMILE	0	There are no Dr. Caglar.
	MLE	3	They also reminded them.

Table 7: Translation examples for min-risk models trained with SIMILE and BLEU and our baseline MLE model.

tence.[14] In contrast, parts-of-speech that in general convey little semantic information like determiners show very little difference in F1 between the two approaches.

6 Qualitative Analysis

We show examples of the output of all three systems in Table 7, along with their human scores which are on a 0-5 scale. The first 5 examples shows cases where SIMILE better captures the semantics than BLEU or MLE. In the first three, the SIMILE model adds a crucial word that the other two systems do not making a significant difference in preserving the semantics of the translation. These words range include verbs (*tells*), prepositions (*For*), adverbs (*viable*) and nouns (*conversation*). The fourth and fifth examples also show how SIMILE can lead to more fluent outputs and is effective on longer sentences.

The last two examples are failure cases of using SIMILE. In the first, it repeats a phrase, just as the MLE model does and is unable to smooth it out as the BLEU model is able to do. In the last example, SIMILE again tries to include words significant to the semantics of the sentence, the entity *Dr. Caglar*. However it misses on the rest of translation, despite being the only system to include this noun phrase.

7 Metric Comparison

We took all outputs of the validation set of the `de-en` data for our best SIMILE and BLEU models as measured by BLEU validation scores and sorted the outputs by the following statistic:

$$|\Delta\text{BLEU}| - |\Delta\text{SIM}|$$

where BLEU refers to sentence-level BLEU. Examples of some of the highest and lowest scoring sentence pairs are shown in Table 8. The top half of the table shows examples where the difference in SIM scores is large, but the difference in BLEU scores is small. From these examples, we see that when SIM scores are different, there can be a difference in how close in meaning the generated sentences are to the reference. When BLEU scores are very close, this may not be the case and it's even possible for less accurate translations to have higher scores than more accurate ones.

The bottom half of the table shows examples where the difference in BLEU scores is large, but the difference in SIM scores is small. From thexe examples we can see that when BLEU scores are very different, the semantics of the sentence can still be preserved. However, we observe that often in these cases, the SIM scores of the sentences tend to be similar.

8 Related Work

The seminal work on training machine translation systems to optimize particular evaluation measures was performed by Och (2003), who intro-

---

[14]Note that in the testing data, INTJ often corresponds to words like *Yes* and *No* which tend to be very important regarding the semantics of the translation in these cases.

Reference	Workers are beginning to clean up workers .
BLEU system	Workers have begun to clean up in Rszke.
SIM system	In Rszke, workers are beginning to clean up.
△BLEU	3.2
△SIM	-26.3
Reference	All that stuff sure does take a toll.
BLEU system	None of this takes a toll .
SIM system	All of this is certain to take its toll .
△BLEU	7.1
△SIM	-22.7
Reference	Another advantage is that they have fewer enemies.
BLEU system	Another benefit: they have less enemies.
SIM system	Another advantage: they have fewer enemies.
△BLEU	-33.8
△SIM	-9.6
Reference	I don't know how to explain - it's really unique.
BLEU system	I do not know how to explain it - it is really unique.
SIM system	I don't know how to explain - it is really unique.
△BLEU	-39.1
△SIM	-2.1

Table 8: The top two rows show examples where the generated sentences have similar BLEU scores but quite different SIM scores. The bottom two rows show the converse. Negative values indicate the SIM system had a higher score for that sentence.

duced minimum error rate training (MERT) and used it to optimize several different metrics in statistical MT (SMT). This was followed by a large number of alternative methods for optimizing machine translation systems based on minimum risk (Smith and Eisner, 2006), maximum margin (Watanabe et al., 2007), or ranking (Hopkins and May, 2011), among many others.

Within the context of SMT, there have also been studies on the stability of particular metrics for optimization. Cer et al. (2010) compared several metrics to optimize for SMT, finding BLEU to be robust as a training metric and finding that the most effective and most stable metrics for training are not necessarily the same as the best metrics for automatic evaluation. The WMT shared tasks included tunable metric tasks in 2011 (Callison-Burch et al., 2011) and again in 2015 (Stanojević et al., 2015) and 2016 (Jawaid et al., 2016). In these tasks, participants submitted metrics to optimize during training or combinations of metrics and optimizers, given a fixed SMT system. The 2011 results showed that nearly all metrics performed similarly to one another. The 2015 and 2016 results showed more variation among metrics, but also found that BLEU was a strong choice overall, echoing the results of Cer et al. (2010). We have shown that our metric stabilizes training for NMT more than BLEU, which is a promising result given the limited success of the broad spectrum of previous attempts to discover easily tunable metrics in the context of SMT.

Some researchers have found success in terms of improved human judgments when training to maximize metrics other than BLEU for SMT. Lo et al. (2013) and Beloucif et al. (2014) trained SMT systems to maximize variants of MEANT, a metric based on semantic roles. Liu et al. (2011) trained systems using TESLA, a family of metrics based on softly matching n-grams using lemmas, WordNet synsets, and part-of-speech tags. We have demonstrated that our metric similarly leads to gains in performance as assessed by human annotators, and our method has an auxiliary advantage of being much simpler than these previous hand-engineered measures.

Shen et al. (2016) explored minimum risk training for NMT, finding that a sentence-level BLEU score led to the best performance even when evaluated under other metrics. These results differ from the usual results obtained for SMT systems, in which tuning to optimize a metric leads to the best performance on that metric (Och, 2003). Edunov et al. (2018) compared structured losses for NMT, also using sentence-level BLEU. They found risk to be an effective and robust choice, so we use risk as well in this paper.

9   Conclusion

We have proposed SIMILE, an alternative to BLEU for use as a reward in minimum risk training. We have found that SIMILE not only outperforms BLEU on automatic evaluations, it correlates better with human judgments as well. Our analysis also shows that using this metric eases optimization and the translations tend to be richer in correct, semantically important words.

This is the first time to our knowledge that a continuous metric of semantic similarity has been proposed for NMT optimization and shown to outperform sentence-level BLEU, and we hope that this can be the starting point for more research in this direction.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei

Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. *Proceedings of SemEval*, pages 497–511.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 32–43.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Meriem Beloucif, Chi-kiu Lo, and Dekai Wu. 2014. Improving MEANT based semantically tuned SMT. In *Proceedings of 11th International Workshop on Spoken Language Translation (IWSLT 2014)*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Association for Computational Linguistics.

Daniel Cer, Christopher D. Manning, and Daniel Jurafsky. 2010. The best lexical metric for phrase-based statistical mt system optimization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 555–563. Association for Computational Linguistics.

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, Copenhagen, Denmark.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, et al. 2018. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of NAACL*, volume 1, pages 355–364.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1352–1362. Association for Computational Linguistics.

Bushra Jawaid, Amir Kamran, Miloš Stanojević, and Ondřej Bojar. 2016. Results of the wmt16 tuning shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 232–238. Association for Computational Linguistics.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the Conference on Computational Linguistics*, pages 501–507.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2011. Better evaluation metrics lead to better machine translation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 375–384, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Chi-kiu Lo, Karteek Addanki, Markus Saers, and Dekai Wu. 2013. Improving machine translation by training against an automatic semantic frame based evaluation metric. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 375–381. Association for Computational Linguistics.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL) Demo Track*, Minneapolis, USA.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*.

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, Philadelphia, Pennsylvania, USA.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2012. On the difficulty of training recurrent neural networks. *arXiv preprint arXiv:1211.5063*.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Nghia The Pham, Germán Kruszewski, Angeliki Lazaridou, and Marco Baroni. 2015. Jointly optimizing word representations for lexical and sentential tasks with the c-phrase model. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremb. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2015. Minimum risk training for neural machine translation. *arXiv preprint arXiv:1512.02433*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692. Association for Computational Linguistics.

David A. Smith and Jason Eisner. 2006. Minimum risk annealing for training log-linear models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 787–794. Association for Computational Linguistics.

Miloš Stanojević, Amir Kamran, and Ondřej Bojar. 2015. Results of the wmt15 tuning shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 274–281. Association for Computational Linguistics.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Taro Watanabe, Jun Suzuki, Hajime Tsukada, and Hideki Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *Proceedings of the International Conference on Learning Representations*.

John Wieting and Kevin Gimpel. 2017. Revisiting recurrent networks for paraphrastic sentence embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2078–2088, Vancouver, Canada.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Simple and effective paraphrastic similarity from parallel translations. In *Proceedings of ACL*.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas. Association for Computational Linguistics.