NUBIA: NeUral Based Interchangeability Assessor for Text Generation

Hassan Kane [*]
MIT
WL Research

Muhammed Yusuf Kocyigit [*]
Boston University
WL Research

Ali Abdalla
WL Research

Pelkins Ajanoh
Harvard University
WL Research

Mohamed Coulibali
Laval University
WL Research

Abstract

We present NUBIA, a methodology to build automatic evaluation metrics for text generation using only machine learning models as core components. A typical NUBIA model is composed of three modules: a neural feature extractor, an aggregator and a calibrator. We demonstrate an implementation of NUBIA showing competitive performance with state-of-the art metrics used to evaluate machine translation and state-of-the art results for image captions quality evaluation. In addition to strong performance, NUBIA models have the advantage of being modular and improve in synergy with advances in text generation models.

1 Introduction

Evaluation metrics play a central role in the machine learning community. They direct research efforts and define the state of-the-art models. Unlike machine learning tasks such as classification and regression, text generation (i.e. machine translation, summarization, image captioning) is a nuanced task where the gold standard for quality evaluation is human assessment. However, this method of evaluation is expensive and time consuming.

As a complement, automatic metrics were designed to approximate human judgment of quality. A consequence of this unique setup is that the metrics themselves have to be frequently upgraded to reflect the dynamic progress of the field. However this has not happened and, while the text generation models have dynamically evolved, the metrics most commonly to assess model outputs used have not.

The two most common metrics used for evaluating similarity between candidate and reference texts are BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004). Both approaches rely on counting the matching n-grams in the candidate text to n-grams in the reference text. The former is precision focused while the latter is recall focused.

These metrics have posed serious limitations and have already been criticized by the academic community (Reiter, 2018; Callison-Burch et al., 2006; Sulem et al., 2018; Novikova et al., 2017). In this work, we propose a methodology to build text generation evaluation metrics using deep learning models as core components.

An implementation of this methodology is then presented and tested in the domains of machine translation and image captioning quality estimation. For assessing the metric in the machine translation domain, we use the WMT 2017, 2018 and 2019 dataset.

We conduct further experiments showing that, without any additional fine-tuning, the same model used to assess machine translation quality outperforms existing metrics specifically designed to assess image captioning quality.

Beyond the promise of this methodology in terms of its ability to lead to metrics with high correlation to human judgment, NUBIA metrics can be constructed with any base architecture, perform well with only thousands of examples as supervision signal and are expected to improve continuously with future NLP advances.

2 Related Work

2.1 BLEU, ROUGE and n-gram matching approaches

BLEU(Papineni et al., 2002) and ROUGE (Lin, 2004) have been used as the main evaluation methods in a variety of NLP tasks for almost two decades. BLEU is shown to better correlate with

[*] Equal contribution. Correspondence to Hassan Kane: <hassanmohamed@alum.mit.edu>

human judgment when the hypothesis texts are bad as we can see in figure 2(c) and correlate weakly when the hypothesis texts are better. CIDEr is an image captioning metric that computes cosine similarity between tf–idf weighted n-grams (Vedantam et al., 2015). METEOR (Banerjee and Lavie, 2005) uses the harmonic mean of unigram precision and recall in combination with synonym matching and stemming along with word matching. While n-gram matching approaches are fast and simple to understand, this paradigm is limited in its ability do capture higher order semantic meaning.

The shortcomings of these methods have been widely criticised and studied. Reiter (2018), in his structured review of BLEU, finds a low correlation between BLEU and human judgment. Callison-Burch et al. (2006) examine BLEU in the context of machine translation and find that BLEU neither correlates with human judgment on adequacy (whether the hypothesis sentence adequately captures the meaning of the reference sentence) nor on fluency (the quality of language in the hypothesis sentence). Sulem et al. (2018) examine BLEU – in the context of text simplification – on grammaticality, meaning preservation and simplicity. They report a very low, and, in some cases, negative correlation with human judgment.

2.2 Transformers, BERT and GPT

Language modeling has become an important NLP technique, thanks its ability to be applied to various NLP tasks as explained in (Radford et al., 2019). There are two leading architectures for language modeling: Recurrent Neural Networks (RNNs) (Mikolov et al., 2010) and Transformers (Vaswani et al., 2017). RNNs handle the input tokens, words or characters, one by one through time and learn the relationship between them, whereas transformers receive a segment of tokens and learn the dependencies between them using an attention mechanism.

The recent success of transformers as multitask learners (Radford et al., 2019) motivated us to adapt them for the task of neural language evaluation. This is crucial because what stood as an obstacle before neural language models was the power to generalize well to different datasets and tasks. Now with models like GPT-2 (Radford et al., 2019) and RoBERTa (Liu et al., 2019) trained on huge amounts of data, we can start trusting their ability to generalize across domains. As of now, machine summarization, translation and image captioning all use different metrics to compare reference sentences with candidate sentences. Transformers-based models offer the promise to unify quality evaluation across tasks.

2.3 Model-based metrics

While BLEU and ROUGE are defined in a discrete space of word tokens, other evaluation metrics are powered by neural networks and word vectors. BERTscore (Zhang* et al., 2020) computes word embeddings and cosine similarity to create a score array and uses greedy matching to maximize the similarity score between words in the candidate and reference sentences. Sentence Mover's Similarity (Clark et al., 2019) uses a Wasserstein metric defined on sentence embeddings generated from averaging the word embeddings in a sentence. YiSi (Lo, 2019) also defines a distance metric among reference and hypothesis sentences based on multilingual BERT embeddings and word frequency weightings. SPICE (Anderson et al., 2016) is an image captioning metric that creates a parse tree from the reference caption, candidate caption to create a scene graph and compute a score based on the overlapping relationships.

These methods report stronger correlations with human judgment and better results when compared with BLEU and ROUGE. While they are using word embeddings (Mikolov et al., 2013) to convert their sentences in a continuous space, they use hand-crafted mathematical functions to evaluate similarity in that space. In NUBIA, rather than defining a mathematical formula, we train a neural network to learn it using human judgement on thousands of sentence pairs as supervision signal.

BLEND (Ma et al., 2017) uses an SVM to combine different existing evaluation metrics. RUSE (Shimanaka et al., 2018) embeds both sentences separately and pools them to a given size. After, the method uses a pre-trained MLP to predict on different tasks. This quality estimator metric is then proposed to be used in language evaluation.

BLEURT (Sellam et al., 2020) introduces a BERT model in combination with a novel pre-training scheme that uses millions of synthetic examples to help the model generalize and then fine-tune it on human judgement.

Our proposed methodology is also a learned metric. Instead of synthesizing millions of examples, we use different pre-trained transformers as feature

extractors on reference, hypothesis sentence pairs and then learn a mapping between those features and a final quality score.

2.4 GLUE Benchmark

The GLUE Benchmark is a collection of tools for evaluating and analyzing the performance of NLP models across a diverse range of tasks (Wang et al., 2018). The recent introduction of this benchmark has encouraged the NLP community to move away from specialized models doing well on a single task to models performing well across diverse tasks. NLP models such as transformers are usually pre-trained on a large corpus in an unsupervised manner and then fine-tuned on a dataset used for the specific task of the benchmark. Architectures doing well on this benchmark can be used as components of future NUBIA models

3 NUBIA model

Our method has three modules: a neural feature extractor, an aggregator and a calibrator. The feature extractor tested in this paper consists of different transformer architectures fine-tuned on relevant tasks of language evaluation such as semantic similarity, logical inference and sentence likelihood. While we use these features and architectures as the main building blocks of NUBIA, the specific models can change as long as they maintain the necessary performance in terms of correlation with human judgment on the fine-tuning tasks.

The aggregator uses the features extracted by the transformers as well as non-neural features such as reference and candidate sentence length and is trained to predict the quality of the hypothesis sentence given the reference sentence. Similar to the WMT challenge, we use past years' data to train this aggregator and test it on the test subset.

The calibrator is the final module that caps all predictions to be between 0 and 1.

3.1 Neural Feature Extraction

In this section, we will describe how we broke down the problem of assessing the quality of a sentence into numerical features, the thought process behind the features used and provide details on the models used for one possible implementation of a NUBIA architecture.

3.1.1 Semantic similarity

The first feature extracted between candidate and reference sentence is semantic similarity. In our proposed implementation, we use a RoBERTa large pre-trained model (Liu et al., 2019), which we fine-tune to predict sentence similarity (0-5 scale) on the STS-B benchmark dataset (8,628 sentence pairs).

The rationale for this feature is that a good candidate sentence should have high semantic similarity with the reference sentence.

3.1.2 Logical Entailment

The second set of features looks at the logical relationship between the reference and hypothesis sentence. The quality of the generated text depends not only on the grammar and semantics but also the core meaning and argument of the candidate sentence. A good model will output sentences that convey the same message.

To extract these features, we use a RoBERTa large pre-trained model (Liu et al., 2019) which is then fine-tuned on the MNLI challenge from the GLUE benchmark.

The MNLI model is trained to take as input sentence pairs and output 0 if the sentences are in contradiction with each other, 1 if the logical relationship is undetermined/neutral (i.e. sentences do not discuss the same topic) and 2 if the sentence are in logical agreement with each other.

We take the likelihood scores over the 3 possible classes as features.

3.1.3 Sentence Intelligibility

The third set of neural features aims to capture the linguistic acceptability of the candidate sentence.

The rationale of this feature is that we want to make sure that candidate sentences are legible and grammatically correct.

It is a common failure mode for machine translation models to generate sentences which are close in meaning to the reference sentence but introduce uncommon syntax and grammatical errors. We currently model this by using the perplexity score of a state-of-the-art Neural Language Model: GPT-2 (Radford et al., 2018)

More precisely, given a sentence A and a sentence B, the 2 features we compute are the perplexity scores for sentence A and sentence B. Optionally, in one of the NUBIA version, we also introduce the number of words in the candidate and reference sentences. We have experimentally found that adding these features in conjunction with the perplexity scores improves correlation with human judgment.

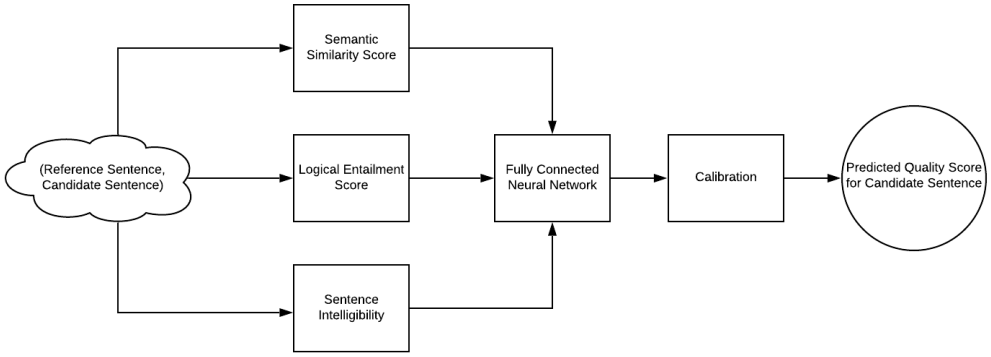


Figure 1: Outline of a NUBIA model with the three steps of Neural Feature Extraction, Aggregation and Calibration.

3.2 Aggregator

In the section above, we defined the dimensions used to assess the quality of a candidate sentence and then showed how to turn these dimensions into numerical scores using transformer models. The aggregator module is trained to approximate a function mapping input neural features to a quality score reflecting how interchangeable the candidate sentence and the reference sentences are.

The inspiration behind this model is that when human evaluators assess the quality of a candidate sentence against a reference sentence, they simultaneously pay attention to several aspects of the candidate sentence such as its semantic similarity with the reference sentence and whether it makes grammatical sense.

Since the relationship between these features and human judgement of quality is unknown, the goal of the aggregator is to approximate it using data obtained from rigorously conducted human evaluations.

The aggregator is a regression model trained to predict human evaluation on pairs of candidate and reference sentences. In this work, we explored linear regression and feed-forward, fully connected neural network architectures.

The neural network aggregator is a fully-connected, feedforward neural network architectures with either 6 (neural features only) or 8 (neural features and number of words in candidate and reference sentences) input layers corresponding to the features extracted, 10 hidden layers and a 1 dimension output layer corresponding to the human score prediction. The activation function for the model is the hyperbolic tangent and the optimizer is ADAM (Kingma and Ba, 2015). NUBIA models using 6 input features have the "NUBIA-6DIM" prefix while the NUBIA models using 8 input features have the "NUBIA-8DIM" prefix. Models using a neural network as an aggregator have the "-NN" suffix while those using linear regression have the "-LREG" suffix.

3.3 Calibration

In practice, the output of the regressors are already highly correlated with human judgement; however, they lack two important properties. The first one is that the regressed score comparing a reference sentence with itself is not always equal to 1. To remedy to this, we normalize the scores given to a candidate sentence by the score given by the regressor of the candidate sentence with itself. The second missing property is that the raw regression scores are not strictly bounded to be between 0 and 1. To ensure they are, we cap the output of the regressors to have a value between 0 and 1.

4 Experiments

To assess our proposed implementation, we used both direct assessment and segment-level relative ranking from different WMT metrics shared tasks (Bojar et al., 2017; Graham et al., 2015) as well as tasks from image captioning. We did not conduct experiments in the domain of machine summarization because there are no labeled datasets containing pairs of summaries and their corresponding human evaluations of the summary quality.

In the WMT Direct Assessment task, candidate and reference translations are given for several language pairs and for each candidate translation, 15 human evaluators assign a quality score between

0 and 100. The final human score is taken as the average of the 15 human assessments. The performance of metrics is assessed using Pearson correlation with human judgement. For this task, we used the 2017 dataset because, unlike the WMT 2018 and WMT 2019 dataset, each sentence has been scored by at least 15 human evaluators (Ma et al., 2018).

For relative ranking, WMT 2018 and WMT 2019 still use direct human assessments but since there is not at least 15 annotators per sentence pairs, the direct assessment correlation task is converted into relative ranking task. More specifically, for a given reference sentences, up to 5 machine translation systems generate candidate translations. These candidate sentences are rated by human annotators on a discrete 0-25-50-75-100 points scale. After averaging the human annotations, if the gap between two candidate translation is higher than 25 points, one translation is considered to be better than the other. When the gap between two candidate sentences is lower than 25 points, the sentence pairs are not included in the segment-level evaluation Ma et al. (2018). In that setting, metrics are scored on their ability to preserve the human ranking using the Kendall's Tau correlation coefficient.

4.1 Model training and testing

4.1.1 Machine Translation

For the machine translation experiments, we use the WMT 2015, 2016, 2017, 2018 and 2019 datasets in different settings. In these datasets, we only picked translations where the target language is English. This was done because the language models we used and their underlying word embeddings are trained on English sentences.All datasets are used for testing in future years.

For the WMT 2017 dataset (3,920 sentence pairs), we use an aggregator trained on human judgement from WMT 2015 and 2016 (5,360 sentence pairs). For the WMT 2018 (207,576 sentence pairs) and WMT 2019 (281,009 sentence pairs), we used an aggregator trained on WMT 2015 through 2017 (9,280 sentence pairs). In practice we found no improvement by adding sentences from WMT 2018 to train the aggregator which is why we stick with WMT 2015 through 2017 to test on both WMT 2018 and WMT 2019.

Feature extraction was conducted using one P100 GPU instance and took 3 hours for WMT 2017 and four days for WMT 2018 and 2019.

For the WMT 2017 task, the performance metric is Pearson correlation with human judgement. For the WMT 2018 and WMT 2019 challenges which are focused on relative ranking, metrics are compared with a Kendall's Tau formulation on how well their scores correlate with human rankings of machine translation.

4.1.2 Image Captioning

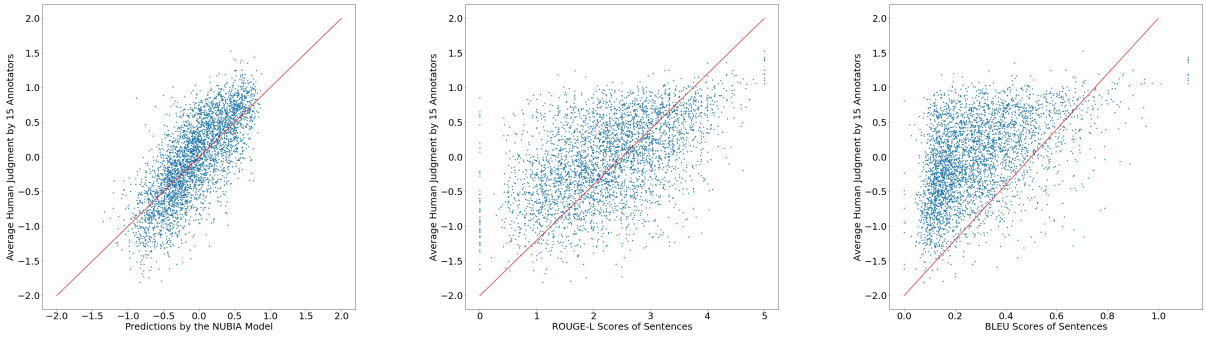
For image captioning, we followed SPICE and used the Flickr 8K dataset. This dataset consists of 8,092 images annotated with 5 gold standard captions generated by humans. The dataset also has a human-evaluated part where for each image, a candidate caption is selected from the entire dataset and scored by three expert judges between 1 ("the selected caption is unrelated to the image") and 4 ("the selected caption describes the image without any error."). This part has 5,822 human-evaluated image caption pairs where each image also has 5 reference gold standard captions.

NUBIA is compared with Kendall's Tau on how well it correlates with the average of the three judges' scores as labels. Neural Feature extraction was conducted using one P100 GPU instance and took 12 hours. The aggregators for the NUBIA models used in the image captioning experiments are not specifically fine-tuned for the task and consist of the Neural Feature Extractors described above along with an aggregator trained on the WMT 2015, WMT 2016 and WMT 2017 dataset (9,280 sentence pairs).

5 Results

In Table 2, we report our results on the test set. We compare our methods with methods developed for the WMT 2017 challenge and recent models like BERTScore and BLEURT which are currently the best performing methods. Although many methods have been proposed throughout the years in the WMT metrics challenge, the current methods used to this day to assess performance of machine translation models are still BLEU and ROUGE score. For ROUGE, we use ROUGE-L scores because it is the formulation of ROUGE correlated the most with human judgements on WMT 2017.

In Table-3, we report the results for the relative ranking test of WMT 2018. Here we see that NUBIA is only outperformed by BLEUR. In Table-4, we have the results for the WMT 2019 challenge. Here we observe that NUBIA performs comparably with other methods.



NUBIA-NN ROUGE-L BLEU

Figure 2: Score and label graphs of NUBIA, ROUGE-L and BLEU for the entire WMT-2017 segment level sets.

Model	τ
BLEU-1*	0.32
BLEU-4	0.33
ROUGE-L*	0.32
BERTScore	0.394
METEOR*	0.42
BLEURT	0.434
CIDEr*	0.44
SPICE*	0.45
NUBIA-6DIM-NN	0.47
NUBIA-8DIM-NN	**0.495**

Table 1: Kendall's Tau Correlation with human judgment on Flickr 8K dataset. The scores marked with * are taken directly from the original SPICE paper. The BLEU-4 score in the original paper was 0.14 but the experiment was repeated with a smoothed function and the new result is reported.

We report the results of the image captioning experiments in Table-1. Here we observe that NUBIA outperforms all existing methods and achieves state-of-the-art correlation with human judgment of caption quality.

The strong performance maintained across varied tasks is a strong indicator of the robustness of this methodology and shows its promise to generalize well beyond the training set.

5.1 Ablation Study

To judge the importance of the features we have picked, we ran an ablation study where we trained a NUBIA model with only a subset of the features and report correlation results on the WMT17 dataset. The most crucial feature is the RoBERTa semantic similarity score. As suspected, other elements beyond semantic similarity also seem to be factored into prediction of translation quality as evidenced by the performance boost obtained after computing the GPT-2 features and MNLI features.

5.2 Error Analysis

Figure 2 sheds more light on the behavior of BLEU and ROUGE, two of the most common evaluation metrics and NUBIA-NN. This analysis unveils important properties of these metrics and helps better understand their strengths and weaknesses.

If we start with (c) we can see that BLEU correlates better with human judgment in the bottom left (bad hypothesis area). Essentially, if a human is likely to give a bad score to a sentence, BLEU is unlikely to overscore. But if a person is going to give a high score, BLEU is equally likely to give any score, maybe even more likely to penalize the sentence. This effectively inhibits the desired behaviour in language generation.

While the behaviour of ROUGE is much more balanced, it is still prone to underscoring and overscoring.

When we look at NUBIA-NN, we see a general trend followed along the data, as expected given the

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	AVG
Human Evaluation	DA	DA	DA	DA	DA	DA	DA	DA
Correlation	r	r	r	r	r	r	r	r
BLEU	0.432	0.425	0.577	0.415	0.479	0.548	0.515	0.484
ROUGE-L	0.482	0.492	0.623	0.465	0.480	0.593	0.569	0.529
BLEND	0.594	0.571	0.733	0.577	0.622	0.671	0.661	0.632
MEANT2.0	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	0.681
NUBIA-6DIM-LReg	0.739	0.733	0.815	0.788	0.734	0.766	0.763	0.763
NUBIA-8DIM-LReg	0.739	0.732	0.829	0.783	0.731	0.784	0.768	0.767
BERTscore	0.714	.740	0.835	0.774	0.773	0.776	0.767	0.768
NUBIA-6DIM-NN	0.745	0.730	0.847	0.779	0.737	0.800	0.751	0.770
NUBIA-8DIM-NN	0.754	.738	0.854	0.786	0.755	0.804	0.750	0.777
BLEURT	**0.773**	**0.792**	**0.878**	**0.835**	**0.811**	**0.824**	**0.814**	**0.818**

Table 2: Absolute Pearson correlations with segment-level human judgments on WMT17 to-English translations. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en	AVG
Human Evaluation	DA	DA	DA	DA	DA	DA	DA	DA
Correlation	τ	τ	τ	τ	τ	τ	τ	τ
BLEU	0.268	0.458	0.311	0.206	0.259	0.178	0.210	0.270
ROUGE-L	0.28	0.473	0.324	0.208	0.275	0.193	0.211	0.281
YiSi1 SRL 18	0.317	0.483	0.345	0.237	0.306	0.233	0.209	0.304
RUSE	0.3478	0.498	0.368	0.273	0.311	0.259	0.218	0.325
YiSi1 SRL 19	0.396	0.543	0.39	0.303	0.351	0.297	0.253	0.362
Yisi1	0.391	0.544	0.397	0.299	0.352	0.301	0.254	0.363
BERTScore	0.408	0.550	0.395	0.293	0.346	0.296	0.260	0.364
NUBIA-6DIM-NN	0.396	0.550	0.410	0.326	0.357	0.295	0.262	0.371
NUBIA-8DIM-NN	0.402	0.553	0.410	**0.330**	0.357	0.288	**0.268**	0.373
BLEURT	**0.423**	**0.567**	**0.414**	0.325	**0.360**	**0.315**	0.260	**0.381**

Table 3: Kendall's Tau correlation with segment-level human judgments on WMT18 to-English translations. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

high correlation score. The only interesting action is the overscoring of low human score sentences. The nature of the error can be analyzed to further improve NUBIA.

6 Conclusion

In this work, we introduced NUBIA: a methodology to build automatic evaluation metrics for text generation using machine learning models as core components. An implementation of this methodology achieves strong results on machine translation and state-of-the art results on image captioning strongly building on the successes of recent NLP architectures such as RoBERTa and GPT-2. These strong results are achieved using a small amount of supervised training data. This methodology offers the possibility of building evaluation metrics improving in synergy with the progress of generative models and unifying evaluation of image captioning, machine translation and potentially other text generation tasks.

7 Discussion and future work

Learned text generation evaluation metrics have enormous promise to change how text generation models are assessed. Future work can further probe which other text generation tasks NUBIA models are strong candidates to assess.

NUBIA can be improved along four axes. The first axis of improvement is through the efforts of the wider NLP community at creating models achieving strong results on the NLU benchmarks like GLUE. The second axis is through the addition of better features capturing aspects of human

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	AVG
Human Evaluation	DA	DA	DA	DA	DA	DA	DA	DA
Correlation	τ	τ	τ	τ	τ	τ	τ	τ
BLEU	0.173	0.264	0.207	0.389	0.280	0.166	0.349	0.261
ROUGE-L	0.169	0.268	0.198	0.394	0.294	0.171	0.348	0.263
ESIM	0.167	0.337	0.303	0.435	0.359	0.201	0.396	0.314
NUBIA-6DIM-NN	0.248	0.356	0.274	0.419	0.385	0.227	0.410	0.331
YISI	0.199	0.346	0.306	0.442	0.380	0.222	0.431	0.332
NUBIA-8DIM-NN	**0.251**	0.358	0.258	0.429	.385	**0.229**	0.413	0.332
BERTscore	0.230	0.345	**0.320**	0.432	0.381	0.223	**0.444**	**0.339**
BLEURT	0.169	**0.363**	**0.319**	**0.446**	**0.406**	0.223	0.424	0.336

Table 4: Kendall's Tau correlation with segment-level human judgments on WMT19 to-English translations. Correlations of metrics not significantly outperformed by any other for that language pair are highlighted in bold.

	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	AVG
Human Evaluation	DA	DA	DA	DA	DA	DA	DA	DA
Correlation	r	r	r	r	r	r	r	r
NUBIA-NN,SI	0.412	0.451	0.624	0.571	0.447	0.437	0.410	0.478
NUBIA-NN,LI	0.620	0.539	0.693	0.647	0.603	0.692	0.571	0.623
NUBIA-NN,LI+SI	0.643	0.621	0.775	0.722	0.646	0.681	0.624	0.673
NUBIA-NN,SS	0.678	0.686	0.790	0.740	0.694	0.766	0.708	0.723
NUBIA-NN,SS+LI	0.696	0.699	0.804	0.758	0.708	0.784	0.723	0.738
NUBIA-NN,SS+SI	0.727	0.729	0.842	**0.785**	0.726	0.790	0.755	0.764
NUBIA-NN,SS+LI+SI	**0.754**	**0.738**	**0.854**	**0.786**	**0.755**	**0.804**	**0.750**	**0.777**

Table 5: Ablation study results for NUBIA-NN on WMT 2017 Direct Assessment task. SS=Semantic Similarity, LI=Linguistic Inference, SI=Sentence Intelligibility.

quality assessment. Two candidate features are the linguistic acceptability which can be obtained by using models trained on the CoLA challenge and a coherence score for long text generations. The third axis is through better aggregator design. Finally, the fourth axis is reducing the computational cost of NUBIA models. The transformer architectures used as backbone for feature extraction are currently independent of each other. Using lighter models or fine-tuning using shared layers could lead to less compute-intensive models.

Learning how to specify NUBIA architectures and standardizing nomenclature will be crucial to ensure adoption, reproducibility and fair comparison of models scored using such automatic metrics. An exhaustive solution can be to describe the individual feature extractor. This description should not only include architectures but also training data and fine-tuning data (Mitchell et al., 2019; Gebru et al., 2018; Bender and Friedman, 2018). Similarly, aggregators should also be described through their architectures along with the training corpus

Evaluation and scorecards for neural metrics going beyond correlation with human judgement (Boag et al., 2016) will help shed lights on their inner workings and failure modes. Such setups more precisely measure the effect that systematic sentence transformations (e.g. active to passive voice) have on the automatic metric scores.

Closely related to evaluation and data reporting, biased training data leading to underscoring or over scoring of valid translations should also be investigated.

Another area of current limitation is the language. Existing NUBIA models only work for English sentence pairs though the procedure to generate and assess such metrics in other languages is likely to be similar.

Understanding how such models can be adversarially attacked is also an open research question.

Finally, future work can also investigate convergence behavior and output of training setups where NUBIA is used as a loss function of text generation models.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

William Boag, Renan Campos, Kate Saenko, and Anna Rumshisky. 2016. MUTT: Metric unit TesTing for language generation tasks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1935–1943, Berlin, Germany. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 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, Colorado. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. *CoRR*, abs/1707.06875.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

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: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.