Learning to Evaluate Translation Beyond English
BLEURT Submissions to the WMT Metrics 2020 Shared Task

Thibault Sellam **Amy Pu*** **Hyung Won Chung†** **Sebastian Gehrmann**
{tsellam, puamy, hwchung, gehrmann}@google.com

Qijun Tan **Markus Freitag** **Dipanjan Das** **Ankur P. Parikh**
{qijuntan, freitag, dipanjand, aparikh}@google.com
Google Research

Abstract

The quality of machine translation systems has dramatically improved over the last decade, and as a result, evaluation has become an increasingly challenging problem. This paper describes our contribution to the WMT 2020 Metrics Shared Task, the main benchmark for automatic evaluation of translation. We make several submissions based on BLEURT, a previously published metric which uses transfer learning. We extend the metric beyond English and evaluate it on 14 language pairs for which fine-tuning data is available, as well as 4 "zero-shot" language pairs, for which we have no labelled examples. Additionally, we focus on English to German and demonstrate how to combine BLEURT's predictions with those of YISI and use alternative reference translations to enhance the performance. Empirical results show that the models achieve competitive results on the WMT Metrics 2019 Shared Task, indicating their promise for the 2020 edition.

1 Introduction

The recent progress in machine translation models has led researchers to question the use of n-gram overlap metrics such as BLEU, which focus solely on surface-level aspects of the generated text, and thus may correlate poorly with human evaluation (Papineni et al., 2002; Lin, 2004; Ma et al., 2019; Mathur et al., 2020; Belz and Reiter, 2006; Callison-Burch et al., 2006). This has led to a surge of interest for more flexible metrics that use machine learning to capture semantic-level information (Celikyilmaz et al., 2020). Popular examples of such metrics include YISI-1 (Lo, 2019), ESIM (Mathur et al., 2019), BERTSCORE (Zhang et al., 2020), the Sentence Mover's Similarity (Zhao et al., 2019; Clark et al., 2019), and BLEURT (Sellam et al., 2020). These metrics utilize contextual embeddings from large models such as BERT (Devlin et al., 2019) which have been shown to capture linguistic information beyond surface-level aspects (Tenney et al., 2019).

The WMT Metrics 2020 Shared Task is the reference benchmark for evaluating these metrics in the context of machine translation. It tests the evaluation of systems that are to-English (X $\rightarrow$ En) and to other languages (X $\rightarrow$ Y), which requires a multilingual approach. An additional challenge for learned metrics is that human ratings are not available for all language pairs, and therefore, the models must use unlabeled data and perform zero-shot generalization.

We describe several learned metrics based on BLEURT (Sellam et al., 2020), originally developed for English data. We first extend BLEURT to the multilingual setup, and show that our approach achieves competitive results on the WMT Metrics 2019 Shared Task.[1] We also present several simple BERT-based baselines, which we submit for analysis. Finally, we focus on English to German and enhance BLEURT's performance by combining its predictions with those of YISI (Lo, 2019) as well as by using alternative references.

2 Background and Notations

Task Reference-based NLG evaluation seeks to assign a score to a triplet of sentences *(input, reference, candidate)*, where *input* is a sentence in the source language, *reference* is a reference translation kept secret at inference time, and *candidate* is a translation produced by an MT system.

[*] Work done during a summer internship. Permanent email address: amy_pu@brown.edu.

[†] Work done as a member of the Google AI Residency Program.

[1] We use the following languages for fine-tuning and/or testing: Chinese, Czech, German, English, Estonian, Finnish, French, Gujarati, Kazakh, Lithuanian, Russian, and Turkish. In addition, we also pre-train on Inuktitut, Japanese, Khmer, Pastho, Polish, Romanian, and Tamil.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 921–927
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

Similar to BLEU (Papineni et al., 2002) and the previous editions of the WMT Metrics shared task, we omit the input and treat the task as a regression problem : we aim to learn a function $\boldsymbol{f} : (\boldsymbol{x}, \tilde{\boldsymbol{x}}) \rightarrow y$ that predicts a quality score y for a candidate sentence $\tilde{\boldsymbol{x}} = (\tilde{x}_1, .., \tilde{x}_p)$ given a reference sentence $\boldsymbol{x} = (x_1, .., x_q)$. The function is supervised on a corpus of N human ratings $\{(\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_i, y_i)\}_{n=1}^{N}$.

BLEURT Most experiments presented in this paper are based on BLEURT, a metric that leverages transfer learning to achieve high accuracy and increase robustness (Sellam et al., 2020). BLEURT is a BERT-based regression model (Devlin et al., 2019). It embeds sentence pairs into a fixed-width vector $\boldsymbol{v}_{\text{BERT}} = \text{BERT}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ with a pre-trained Transformer, and feeds this vector to a linear layer:

$$\hat{y} = \boldsymbol{f}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \boldsymbol{W}\boldsymbol{v}_{\text{BERT}} + \boldsymbol{b}$$

where $\boldsymbol{W}$ and $\boldsymbol{b}$ are the weight matrix and bias vector respectively.

In its original (English) version, BLEURT is trained in three stages. (1) It is initialized from a publicly available BERT checkpoint. (2) The model is then "warmed up" by exposing it to millions of sentence pairs $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$, obtained by randomly perturbing sentences from Wikipedia. During this phase, the model learns to predict a wide range of similarity scores that include existing metrics (BERTSCORE, BLEU, ROUGE), scores from an entailment model, and the likelihood that $\tilde{\boldsymbol{x}}$ was generated from $\boldsymbol{x}$ with a round-trip translation by a given translation model. We denote this stage as *mid-training*. (3) In the final stage, the model is fine-tuned on human ratings from WMT Metrics (Bojar et al., 2017; Ma et al., 2018, 2019), using a regression loss $\ell_{\text{supervised}} = \frac{1}{N}\sum_{n=1}^{N} \|y_i - \hat{y}\|^2$. We found that English BLEURT achieved competitive performance on four academic datasets, WebNLG (Gardent et al., 2017), and the WMT Metrics Shared Task years 2017 to 2019.

3 Extending BLEURT Beyond English

3.1 Modeling

An approach to extend BLEURT would be to use MBERT, the public version of BERT pre-trained on 104 languages, and "mid-train" with non-English signals as described above. Yet, the evidence we gathered from early experiments were inconclusive. On the other hand, we did observe that models trained on several languages were often more accurate than monolingual models, possibly due to the larger amount of fine-tuning data. Thus, we opted for a simpler approach where we start with a multilingual BERT model and fine-tune it on all the human ratings data available for all languages (`X` $\rightarrow$ `Y` and `X` $\rightarrow$ `En`). In most cases, we found that such models could perform zero-shot evaluation: if a language `Y` does not have human ratings data, the metric can still perform evaluation in this target language as long as the base multilingual BERT model contains unlabeled data for `Y`, as observed in the past literature (Karthikeyan et al., 2019; Pires et al., 2019).

We experiment with two pre-trained multilingual models: MBERT and MBERT-WMT, a custom multilingual variant of BERT. The MBERT-WMT model is larger that MBERT (24 Transformer layers instead of 12), and it was pre-trained on 19 languages of the WMT Metrics shared task 2015 to 2020.

Details of MBERT-WMT pre-training We trained MBERT-WMT model with an MLM loss (Devlin et al., 2019), using a combination of public datasets: Wikipedia, the WMT 2019 News Crawl (Barrault et al.), the C4 variant of Common Crawl (Raffel et al., 2020), OPUS (Tiedemann, 2012), Nunavut Hansard (Joanis et al., 2020), WikiTitles[2], and ParaCrawl (Esplà-Gomis et al., 2019). We trained a new WordPiece vocabulary (Schuster and Nakajima, 2012; Wu et al., 2016), since the original vocabulary of mBERT does not support the alphabets of Pashto, Khmer and Inuktitut. The model was trained for 1 million steps with the LAMB optimizer (You et al., 2020), using the learning rate 0.0018 and batch size 4096 on 64 TPU v3 chips.

3.2 Experimental Setup

Datasets At the time of writing, no human ratings data is available for WMT Metrics 2020. Therefore, we use the human ratings from WMT Metrics years 2015 to 2019 for both training and evaluation. We do so in two stages. In the first stage, we use 2015 to 2018 for training (216,541 sentence pairs in 8 languages), setting 10% aside for early stopping. We use 2019 as a development set, to choose hyper-parameters and to

[2] https://linguatools.org/tools/corpora/wikipedia-parallel-titles-corpora/

support high-level modeling decisions. In the second stage, we use 2015 to 2019, that is, all the data available, for training and uniformly sample 10% of the data for early stopping and hyper-parameter tuning. This adds 289,895 sentence pairs and 4 additional languages to our training set, approximately doubling the size of the training data. We report our results on the first setup, but submit our predictions to the shared task using the second setup.

Hyper-parameters We run grid search on the learning rate and export the best model, using values {5e-6, 8e-6, 9e-6, 1e-5, 2e-5, 3e-5}. We use batch size 32 and evaluate the model every 1,000 steps on a 10% held-out data set to prevent over-fitting. During preliminary experiments, we additionally experimented with the batch size, dropout rate, frequency of continuous evaluation, balance of languages, pre-training schemes, WordPiece vocabularies, and model architecture.

3.3 Additional Models and Baselines

English BLEURT We fine-tune a new BLEURT checkpoint, following the methodology described above. The main difference with Sellam et al. (2020) is that we incorporate the to-English ratings of year 2019, which were not previously available.

Monolingual baselines based on BERT We experiment with three baselines and submit the results to the WMT Metrics Shared Task for analysis. BERT-L2-BASE and BERT-L2-LARGE are two regression models based on BERT and trained on to-English ratings. We use the same setup as English BLEURT, but we omit the mid-training phase. A similar approach was described in Shimanaka et al. (2019). BERT-CHINESE-L2 is similar to BERT-L2-BASE, but it uses BERT-CHINESE and it is fine-tuned on to-Chinese ratings.

Other Systems We compare our setups to other state-of-the-art learned metrics: BERTSCORE (Zhang et al., 2020), and Yisi (Lo, 2019) all apply rules on top of BERT embeddings while ESIM (Mathur et al., 2019) is a neural sentence similarity model. PRISM (Thompson and Post, 2020) trains a multilingual translation model that is used as a zero-shot paraphrasing system. All the aforementioned systems take sentences pairs as input. Concurrent work has investigated incorporating the source with great success (Rei et al., 2020). We leave this line of research for future work.

4 Results

Tables 1 and 2 show the results in the X → En direction, at the segment- and system-level respectively. In the majority of cases, one of the BLEURT configurations yields the strongest results. The original BLEURT metric seems to perform better at the segment-level. At the system-level it may be dominated by PRISM (3 out of 7 language pairs) or by one of the simpler BERT-based models (4 out of 7 language pairs).

Tables 3 and 4 present the results for the other languages. MBERT-WMT yields solid results at the segment-level (it achieves the highest correlations for 7 out of 11 language pairs), in particular for the "zero-shot" setups, En → Gu, En → Kk, and En → Lt. It outperforms MBERT consistently, except for En → Ru and En → Zh where it lags behind the other metrics. The results are consistent at the system-level.

Strategy for the WMT Metrics Shared Task Based on these results, we make two "competitive" submissions. We present BLEURT as described above, which we ran on all the X → En sentence pairs. Additionally, we submitted a multilingual system that combines MBERT-WMT (for all languages except Chinese) and BERT-CHINESE-L2 (for Chinese). We ran the multilingual system for all language pairs including to-English, as the large amount of non-English fine-tuning data made available in 2019 may benefit this setup too. We also release the predictions of BERT-BASE-L2, BERT-LARGE-L2, and MBERT for analysis.

5 Additional Improvements on English→German

For English→German, the organizers of WMT20 provide three different reference translations: two standard references and one additional paraphrased reference. Given this novel setup, we investigate how to combine our predictions. Moreover, we use a similar framework to ensemble the predictions of different metrics. In particular, we average the predictions of BLEURT, YISI-1 and YISI-2. All three metrics are different in their approaches. While BLEURT and YISI-1 are reference-based metrics, YISI-2 is reference-

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	**avg**
YiSi	0.164	0.347	0.312	0.440	0.376	0.217	0.426	0.326
YiSi1-SRL	0.199	0.346	0.306	0.442	0.380	0.222	0.431	0.332
ESIM	0.167	0.337	0.303	0.435	0.359	0.201	0.396	0.314
BERTscore	0.176	0.345	**0.320**	0.432	0.381	0.223	0.430	0.330
PRISM	**0.204**	0.357	0.313	0.434	0.382	0.225	0.438	0.336
BLEURT Configurations, English-only								
BERT-L2-base	0.142	0.326	0.274	0.406	0.367	0.197	0.358	0.296
BERT-L2-large	0.172	0.361	0.305	0.424	0.388	0.210	0.420	0.326
BLEURT	0.175	**0.365**	0.316	**0.451**	0.397	0.223	**0.444**	**0.339**
BLEURT Configurations, Multi-lingual								
mBERT	0.172	0.352	0.300	0.430	0.388	0.222	0.397	0.323
mBERT-WMT	0.187	0.363	0.306	0.439	**0.398**	**0.226**	0.425	0.335

Table 1: Segment-level agreement with human ratings on the WMT19 Metrics Shared Task on the to-English language pairs. The metric is WMT's Direct Assessment metric, a robust variant of Kendall τ. The scores for YiSi, YiSi1-SRL, and ESIM come from Ma et al. (2019). The scores for BERTscore and PRISM come from Thompson and Post (2020).

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	**avg**
YiSi	0.949	0.989	0.924	0.994	0.981	0.979	0.979	0.971
YiSi1-SRL	0.950	0.989	0.918	0.994	0.983	0.978	0.977	0.969
ESIM	0.941	0.971	0.885	0.986	0.989	0.968	0.988	0.961
BERTscore	0.949	0.987	0.981	0.980	0.962	0.921	0.983	0.966
PRISM	**0.954**	0.983	0.764	**0.998**	**0.995**	0.914	0.992	0.943
BLEURT Configurations, English-only								
BERT-L2-base	0.938	**0.992**	**0.930**	0.992	0.991	0.976	**0.997**	**0.974**
BERT-L2-large	0.940	0.987	0.819	0.992	0.990	**0.985**	0.993	0.958
BLEURT	0.943	0.989	0.865	0.996	**0.995**	0.984	0.990	0.966
BLEURT Configurations, Multi-lingual								
mBERT	0.937	0.976	0.863	0.984	0.978	0.959	0.978	0.954
mBERT-WMT	0.950	0.991	0.815	0.989	0.992	0.968	0.980	0.955

Table 2: System-level agreement with human ratings on the WMT19 Metrics Shared Task on the to-English language pairs. The metric is Pearson's correlation. The scores for YiSi, YiSi1-SRL, and ESIM come from Ma et al. (2019). The scores for BERTscore and PRISM come from Thompson and Post (2020).

free and calculates its score by comparing translations only to the source sentence. BLEURT is fine-tuned on previous human ratings, while YiSi-1 is based on the cosine similarity between BERT embeddings of the reference and the candidate.

In the remainder of this section, we report BLEURT results using the mBERT-WMT setup unless specified otherwise.[3]

5.1 Modifications to YiSi-1

Before combining BLEURT and YiSi, we perform a series of modifications to YiSi-1 and evaluate their impact on English→German.

Experimental Setup All experimental results are summarized in Table 5. We report both segment-level (DARR) and system-level (Kendall τ) correlations. To replicate the multi-reference setup of 2020, we compute correlations with the standard WMT references as well as the paraphrased reference from Freitag et al. (2020).

Improving YiSi's Predictions Our baseline is similar to the YiSi-1 submission from WMT 2019 (Lo, 2019): we run YiSi-1 with the public multilingual mBERT checkpoint. We then experiment with the underlying checkpoint. We continued pre-training mBERT on the in-domain German NewsCrawl dataset. The resulting model *+pre-train NewsCrawl layer 9* increases the correlation for both reference translations. We improve the correlation further on the paraphrased reference by using the 8th instead of the 9th layer.

Other experiments We tried pre-training BERT on forward translated sentences from German NewsCrawl, to adapt the word embeddings to MT outputs. We also trained a BERT model from scratch on the German NewsCrawl data. These experiments did not result in higher correlations with human ratings.

[3]We use a different checkpoint from the one described in Section 4. The model was trained for 880K steps instead of 1 million, and it uses a sequence length of 256 tokens instead of 128.

	en-cs	en-de	en-fi	*en-gu*	*en-kk*	*en-lt*	en-ru	en-zh	de-cs	*de-fr*	fr-de	**avg**
YiSi1	0.475	0.351	0.537	0.551	0.546	0.470	0.585	0.355	0.376	0.349	0.310	0.446
YiSi1-SRL	-	0.368	-	-	-	-	-	0.361	-	-	0.299	-
ESIM	-	0.329	0.511	-	0.510	0.428	0.572	0.339	0.331	0.290	0.289	-
BERTSCORE	0.485	0.345	0.524	0.558	0.533	0.463	0.580	0.347	0.352	0.325	0.274	0.435
PRISM	0.582	**0.426**	0.591	0.313	0.531	0.558	0.584	0.376	0.458	**0.453**	0.426	0.482
BLEURT Configurations												
BERT-CHINESE-L2	-	-	-	-	-	-	-	0.356	-	-	-	-
MBERT	0.506	0.364	0.551	0.550	0.529	0.516	**0.592**	**0.381**	0.385	0.388	0.291	0.459
MBERT-WMT	**0.603**	0.422	**0.615**	**0.577**	**0.558**	**0.584**	0.492	0.337	**0.461**	0.449	**0.427**	**0.502**

Table 3: Segment-level agreement with human ratings on the WMT19 Metrics Shared Task on non-English language pairs. The metric is WMT's Direct Assessment metric, a robust variant of Kendall τ. Languages without fine-tuning data are denoted in *italics*. The scores for YiSi, YiSi1-SRL, and ESIM come from Ma et al. (2019). The scores for BERTSCORE and PRISM come from Thompson and Post (2020).

	en-cs	en-de	en-fi	*en-gu*	*en-kk*	*en-lt*	en-ru	en-zh	de-cs	*de-fr*	fr-de	**avg**
YiSi1	0.962	**0.991**	0.971	0.909	0.985	0.963	**0.992**	0.951	0.973	0.969	0.908	0.961
YiSi1-SRL	-	**0.991**	-	-	-	-	-	0.948	-	-	0.912	-
ESIM	-	**0.991**	0.957	-	0.980	**0.989**	0.989	0.931	0.980	0.950	0.942	-
BERTSCORE	0.981	0.990	0.970	0.922	0.981	0.978	0.989	0.925	0.969	0.971	0.899	0.961
PRISM	0.958	0.988	0.949	0.624	0.978	0.937	0.918	0.898	0.976	0.936	0.911	0.916
BLEURT Configurations												
BERT-CHINESE-L2	-	-	-	-	-	-	-	**0.953**	-	-	-	-
MBERT	0.942	0.987	0.953	0.949	0.982	0.950	0.947	0.949	0.972	0.970	0.924	0.957
MBERT-WMT	**0.993**	**0.991**	**0.987**	**0.959**	**0.993**	**0.989**	0.888	**0.953**	**0.986**	**0.988**	**0.962**	**0.972**

Table 4: System-level agreement with human ratings on the WMT19 Metrics Shared Task on non-English language pairs. The metric is Pearson's correlation. Languages without finetuning data are denoted in *italics*. The scores for YiSi, YiSi1-SRL, and ESIM come from Ma et al. (2019). The scores for BERTSCORE and PRISM come from Thompson and Post (2020).

Ref	Metric	model	sys-level Kendall τ	seg-level DARR
std	BLEURT	MBERT-WMT¶	**0.896**	**0.420**
std	YiSi-1	MBERT *(WMT19 subm.)*	0.810	0.351
		+pre-train NewsCrawl layer 9	0.870	0.373
		+pre-train NewsCrawl layer 8 †	0.853	0.376
para	BLEURT	MBERT-WMT¶	0.852	**0.413**
para	YiSi-1	MBERT *(WMT19 subm.)*	0.844	0.316
		+pre-train NewsCrawl layer 9	0.887	0.365
		+pre-train NewsCrawl layer 8 †	**0.896**	0.373
src	YiSi-2	MBERT¶	0.307	0.106
2std+para	YiSi-comb	comb of 3 († systems)	**0.905**	0.399
	all-comb	avg of 7 († & ¶ systems)	0.878	**0.454**

Table 5: Agreement with human ratings on the WMT19 Metrics Shared Task for English→German. The first set of results are generated by using the standard reference translations for WMT 2019. The second set of results is generated by using the paraphrased reference translations. YiSi-2 is reference free and only uses the source sentences.

5.2 Combining BLEURT, YiSi-1 and YiSi-2 on Multiple References

We describe our two submissions to WMT 2020, YiSi-COMB and ALL-COMB, which result from our efforts to use multiple references for automatic evaluation. YiSi-COMB is a multi-reference version of the YiSi score (Lo, 2019) aimed at achieving better system-level correlations. ALL-COMB leverages metrics from BLEURT, YiSi-1, and YiSi-2 on multiple references to achieve better segment-level correlation.

YiSi-COMB YiSi scores are F_1 scores of YiSi precision and YiSi recall. For the YiSi-COMB submission, we take the minimum of the YiSi recalls for the three different references as the multi-reference recall, and the maximum of the YiSi precision as the multi-reference precision. Using the same notations as in (Lo, 2019), the final score is the F_1 of the recall and precision computed with $\alpha = 0.7$ (see Figure 1). This submission aims to maximize the system-level correlation.

As shown in Table 5, YiSi-1 has the highest system-level correlation on paraphrased references. Given that we used $\alpha = 0.7$, YiSi scores are quite similar to YiSi recalls (when $\alpha = 1.0$, YiSi scores are equal to YiSi recalls). YiSi-1 scores for paraphrased references are usually much lower than those of standard references, therefore taking the minimum recall is oftentimes equivalent to taking the YiSi recall from the paraphrased references. Furthermore, we found that using the maximum precision, in combination with aggregating recalls, usually performs the best.

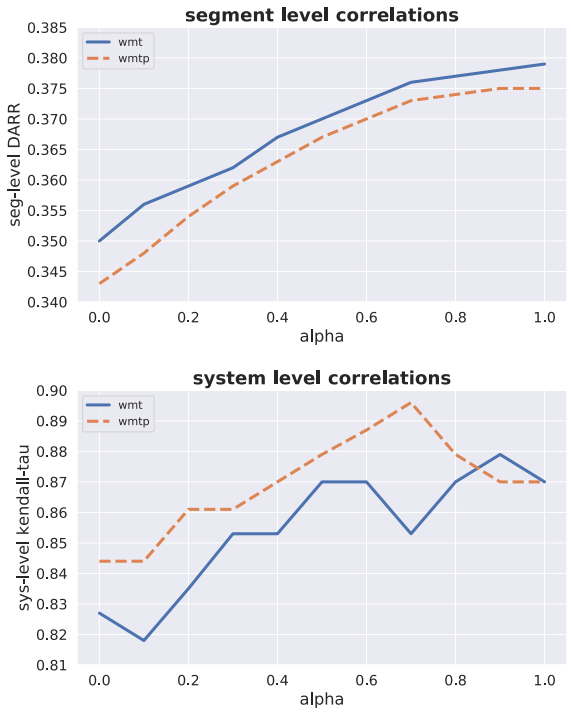

Figure 1: Correlations with respect to different α settings for Yisi-1. The system-level correlation is highest when $\alpha = 0.7$, which is the α we use for the submission.

ALL-COMB We combined the predictions of YISI-1 with those of BLEURT and YISI-2. YISI-2 usually performs worse than the reference-based metrics, but we found that incorporating its predictions can help. Having three different metrics (BLEURT, YISI-1, YISI-2) and three different reference translations, we take all seven predictions and average the scores for each segment. The combined prediction ALL-COMB outperforms every single metric at the segment level, though the system-level correlation drops in comparison to the best YISI-1 score on paraphrased references. This submission aims to maximize the segment-level correlation.

6 Summary

We submit the following systems to the WMT Metrics shared task:

- BLEURT as previously published, fine-tuned on the human ratings of the WMT Metrics shared task 2015 to 2019, to-English.
- A multi-lingual extensions of BLEURT based on a 20 languages variant of MBERT and BERT-CHINESE.
- Three baseline systems based on BERT-BASE, BERT-LARGE, and MBERT.
- Two combination methods for English to German that use YiSi and alternative references, YISI-COMB and ALL-COMB.

7 Acknowledgements

Thanks to Xavier Garcia and Ran Tian for advice and proof-reading.

References

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. Findings of the 2019 conference on machine translation. In *Proceedings of WMT*.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *Proceedings of EACL*.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the wmt17 metrics shared task. In *Proceedings of WMT*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *Proceedings of EACL*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv*.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL HLT*.

Miquel Esplà-Gomis, Mikel L Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. Paracrawl: Web-scale parallel corpora for the languages of the eu. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be Guilty but References are not Innocent. In *Proceedings of EMNLP*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from rdf data. In *Proceedings of INLG*.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The nunavut hansard inuktitut–english parallel corpus 3.0 with preliminary machine translation results.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. In *Proceedings of ICLR*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out*.

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 WMT*.

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 WMT*.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. In *Proceedings of WMT*.

Nitika Mathur, Tim Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *Proceedings of ACL*.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Journal of Machine Learning Research*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv*.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *Proceedings of ICASSP*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *Proceedings of ACL*.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine translation evaluation with bert regressor. *arXiv*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of ACL*.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. *Proceedings of EMNLP*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of The 8th Language Resources and Evaluation Conference*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv*.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training bert in 76 minutes. In *Proceedings of ICLR*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Proceedings of ICLR*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *Proceedings of EMNLP*.