Generative Pretraining for Paraphrase Evaluation

Jack Weston, Raphaël Lenain, Udeepa Meepegama, Emil Fristed
Novoic
{jack, raphael, udeepa, emil}@novoic.com

Abstract

We introduce ParaBLEU, a paraphrase representation learning model and evaluation metric for text generation. Unlike previous approaches, ParaBLEU learns to understand paraphrasis using generative conditioning as a pretraining objective. ParaBLEU correlates more strongly with human judgements than existing metrics, obtaining new state-of-the-art results on the 2017 WMT Metrics Shared Task. We show that our model is robust to data scarcity, exceeding previous state-of-the-art performance using only 50% of the available training data and surpassing BLEU, ROUGE and METEOR with only 40 labelled examples. Finally, we demonstrate that ParaBLEU can be used to conditionally generate novel paraphrases from a single demonstration, which we use to confirm our hypothesis that it learns abstract, generalized paraphrase representations.

1 Introduction

Representing the relationship between two pieces of text, be it through a simple algorithm or a deep neural network, has a long history and diverse use-cases that include the evaluation of text generation models (Wiseman et al., 2017; Van Der Lee et al., 2019) and the clinical evaluation of human speech (Johnson et al., 2003; Weintraub et al., 2018). One of the earliest examples of such a representation is the Levenshtein distance (Levenshtein, 1966), which describes the number of character-level edits required to transform one piece of text into another. This metric now forms part of a wider family of edit-distance-based metrics that includes the word error rate (WER) and the translation error rate (TER) (Och, 2003). Other algorithms, such as ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and the widely used BLEU metric (Papineni et al., 2002), perform exact or approximate n-gram matching between the two texts.

These low-level approaches bear little resemblance to the human process of comparing two texts, which benefits from a deep prior understanding of the semantic and syntactic symmetries of language (Novikova et al., 2017). For example, pairs like "she was no ordinary burglar" and "she was an ordinary burglar" are close in edit-distance-space but semantically disparate. The goal of an automatic text evaluation metric is typically to be a good proxy for human judgements, which is clearly task-dependent. More recently, neural approaches have begun to close the gap between automatic and human judgements of semantic text similarity using Transformer-based language models such as BERT (Zhang et al., 2019a; Sellam et al., 2020). They aim to leverage the transferable knowledge gained by the model during pretraining on large text corpora. The relationship between two texts is similarly modelled, albeit implicitly, by sequence-to-sequence models such as BART (Lewis et al., 2019) and T5 (Raffel et al., 2019). We consider paraphrase evaluation and paraphrase generation to be two instances of *paraphrase representation learning*.

Linguistically, a paraphrase is a restatement that preserves essential meaning, with arbitrary levels of literality, fidelity and completeness. In practice, what qualifies as a good paraphrase is context-specific. One motivation for considering paraphrase evaluation as a representation learning problem is the varied nature of paraphrase evaluation tasks, which may have an emphasis on semantic equivalence (e.g. PAWS (Zhang et al., 2019b) and MRPC (Dolan and Brockett, 2005)), logical entailment versus contradiction (e.g. MultiNLI (Williams et al., 2017) and SNLI (Bowman et al., 2015)), and the acceptability of the generated text (e.g. the WMT Metrics Shared Task (Bojar et al., 2017)). Considering even broader applications such as clinical speech analysis further motivates learning generalized paraphrase representations.

In this paper, we introduce ParaBLEU, a paraphrase representation learning model that predicts a

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4052 - 4073

conditioning factor for sequence-to-sequence paraphrase generation as one of its pretraining objectives, inspired by style transfer in text-to-speech (Skerry-Ryan et al., 2018) and text generation systems (Yang et al., 2018; Lample et al., 2018). ParaBLEU addresses the primary issue with neural paraphrase evaluation models to date: the selection of a sufficiently generalized pretraining objective that primes the model for strong performance on downstream paraphrase evaluation tasks when data is scarce. Previous state-of-the-art neural models have either used a broad multi-task learning approach or eschewed additional pretraining altogether. The former case may encourage the model to learn the biases of inferior or inappropriate metrics, while the latter leaves room for optimization. Non-neural models, such as BLEU, TER, ROUGE and BERTScore (Zhang et al., 2019a), benefit from requiring no training data and thereby avoid domain shift issues. They cannot, however, learn to exploit task-specific nuances of what defines 'good' paraphrasing.

We evaluate ParaBLEU's ability to predict human judgements of paraphrases using the English subset of the 2017 WMT Metrics Shared Task. A useful neural text similarity metric should be robust to data scarcity, so we assess performance as a function of the fine-tuning dataset size. Finally, using the ParaBLEU pretraining model as a paraphrase generation system, we explore our hypothesis that the model reasons in high-level paraphrastic concepts rather than low-level edits through an explainability study, and demonstrate that ParaBLEU can operate as a conditional paraphrase generation model.

2 Approach

In this section, we describe and justify the set of inductive biases we build into ParaBLEU, along with a description of the model architecture and pretraining/fine-tuning strategy. We consider a reference text x and a candidate text $\hat{x}$. We wish to learn a function $f : f(x, \hat{x}) \rightarrow y$, where $y \in \mathbb{R}^{\mathbb{N}}$ is a single- or multi-dimensional paraphrase representation, which could be a scalar score.

2.1 Inductive biases

Our approach begins by decomposing paraphrase representation learning into three overlapping factors:

1. **Edit-space representation learning:** Building a representation of high-level syntactic and semantic differences between x and $\hat{x}$, contrasted with the low-level pseudo-syntactic/-semantic operations considered by edit-distance-based and n-gram based metrics.

2. **Candidate acceptability judgement:** Evaluating the grammaticality, coherence and naturalness of $\hat{x}$ in isolation. Perplexity (Jelinek et al., 1977) with respect to a given language model is one proxy for this.

3. **Semantic equivalence:** Assessing whether x and $\hat{x}$ convey the same essential meaning precisely, as opposed to merely being semantically similar. This is related to entailment classification tasks and, more broadly, the interaction between language and formal logic.

Exploiting this factorization, we hypothesize that the following inductive biases are beneficial to a paraphrase representation learning model:

- **Using pretrained language models:** All three factors require a general understanding of the semantic and syntactic structures of language, making transfer learning from powerful pretrained language models such as BERT (Devlin et al., 2018) appealing.

- **Non-local attention as bitext alignment:** Factors (1) and (3) require performing context-aware 'matching' between x and $\hat{x}$. This is similar to the statistical method of bitext alignment (Tiedemann, 2011). Attention mechanisms within a Transformer (Vaswani et al., 2017) are an obvious candidate for learnable context-aware matching, which has precedent in paraphrasing tasks and the next-sentence-prediction objective of the original BERT pretraining. If the tokens of x and $\hat{x}$ are concatenated into one long input sequence, local attention mechanisms, such as those used in T5, may be suboptimal for longer text-pairs.

- **Bottlenecked conditional generation objective:** A key insight is that a strong factor (1) representation $z \in \mathbb{R}^M$ where $h : h(x, \hat{x}) \rightarrow z$ is one that can condition the sampling of $\hat{x}$ from x through some generative model $g : g(x \mid z) \rightarrow \hat{x}$. One trivial solution to this is $h(x, \hat{x}) = \hat{x}$. To avoid this case, we introduce a bottleneck on z such that

it is advantageous for the model to learn to represent high-level abstractions, which are cheaper than copying $\hat{x}$ through the bottleneck. It is likely advantageous to use a pretrained sequence-to-sequence language model, which can already reason in linguistic concepts.

- **Masked language modelling objective:** Factor (2) can be addressed by an MLM objective, which alone is sufficient for a neural network to learn a language model (Devlin et al., 2018). Performing masked language modelling on a reference-candidate pair also encourages the network to use x to help unmask $\hat{x}$ and vice versa, strengthening the alignment bias useful for factors (1) and (2).

- **Entailment classification objective:** Factor (3) is similar to the classification of whether x logically entails $\hat{x}$. There are a number of sentence-pair datasets with entailment labels that could be used to construct this loss; see Table 4.

2.2 ParaBLEU

Inspired by style transfer in text-to-speech (Skerry-Ryan et al., 2018) and text generation systems (Yang et al., 2018; Lample et al., 2018), we propose the architecture shown in Figure 1. The grey box indicates the Transformer encoder we wish to pretrain, which we refer to as the 'edit encoder'. Factorization of the task leads to three complementary objectives: a cross-entropy masked language modelling loss $\mathcal{L}_{\text{MLM}}$ (Devlin et al., 2018), a cross-entropy autoregressive causal language modelling loss $\mathcal{L}_{\text{AR}}$ (Radford et al., 2018) and a binary cross-entropy entailment classification loss $\mathcal{L}_{\text{CLS}}$. An additional sequence-to-sequence Transformer model is used during pretraining to provide a learning signal. The proposed bottleneck lies within the feedforward network module (see Figure 1), implemented by restricting the hidden dimension to 64 (down from 768 or $1,024$ in the cases of ParaBLEU$_{\text{base}}$ and ParaBLEU$_{\text{large}}$ respectively) before projecting back up to the dimension of the BART decoder. The full pretraining loss is given by:

$$\mathcal{L}_{\text{pre}} := \mathcal{L}_{\text{AR}} + \alpha \cdot \mathcal{L}_{\text{MLM}} + \beta \cdot \mathcal{L}_{\text{CLS}}, \quad (1)$$

where α and β are tunable hyperparameters. We probe the importance of each objective in the ablation studies in Section 4.2. At fine-tuning time, the sequence-to-sequence model is discarded and the edit encoder is fine-tuned using a linear projection on top of the pooled output, projecting the pooled output down to a single dimension that constitutes the predicted score. Throughout this work, our pooling layers simply take the beginning-of-sequence token. An MSE loss $\mathcal{L}_{\text{MSE}}$ is used during fine-tuning.

Our architecture places restrictions on valid combinations of pretrained models. We found in practice that using an encoder-only pretrained language model to initialize the edit encoder, and a sequence-to-sequence pretrained language model to initialize the sequence-to-sequence model, works best. This is likely because encoder-only models are encouraged to encode strong representations at the final layer, and these representations have already been directly pretrained with an MLM objective. For technical ease we require that the models use the same tokenizer, and that the pretrained checkpoints are available through the HuggingFace `transformers` library (Wolf et al., 2019). In this paper, we consider the combination RoBERTa (Liu et al., 2019) + BART, but we note that both multilingual (XLM-R (Conneau et al., 2019) + mBART (Liu et al., 2020)) and long (Longformer + Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020)) combinations exist. We consider both base and large variants, which correspond to RoBERTa$_{\text{base}}$ and RoBERTa$_{\text{large}}$. In both cases, we use a BART$_{\text{base}}$ checkpoint.

2.3 Related work

Evaluation metrics BERTScore (Zhang et al., 2019a), a non-learned neural metric, uses a matching algorithm on top of contextualized neural word embeddings, similar to n-gram matching approaches. MoverScore (Zhao et al., 2019) is similar to BERTScore but uses an optimal transport algorithm. BLEU, ROUGE, METEOR and chrF++ (Popović, 2017) are widely used n-gram-based methods, working at the word, subword or character level. TER is an edit-distance-based metric, similar to WER. BLEURT (Sellam et al., 2020) is a neural automatic evaluation metric for text generation. Starting from a pretrained BERT model, it is further pretrained to predict a number of pre-existing metrics, such as BLEU, ROUGE and BERTScore. ParaBLEU, by contrast, does not use pre-existing metrics as training objectives, instead using generative conditioning as a more

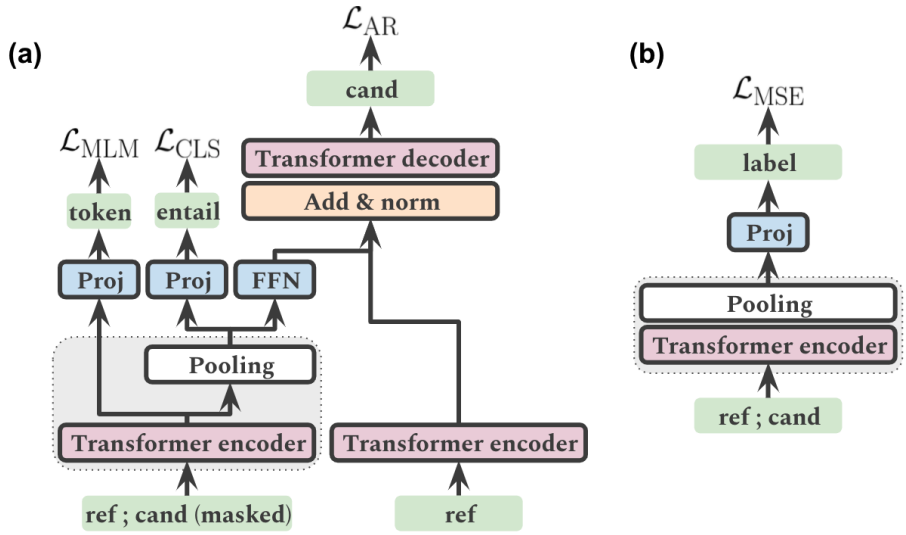

Figure 1: (a) The pretraining setup. (b) The fine-tuning setup. '`ref ; cand`' indicates the canonical method for combining a reference and candidate sentence for a given language model. $\mathcal{L}_{\text{AR}}$ is an autoregressive causal language modelling loss, $\mathcal{L}_{\text{MLM}}$ a masked language modelling loss, and $\mathcal{L}_{\text{CLS}}$ an entailment classification loss. The fine-tuning loss ($\mathcal{L}_{\text{MSE}}$) is a mean-squared error loss. The feedforward network (FFN) includes two affine layers, the middle dimension of which can be used to create a bottleneck (see Section 2.1). Dropout layers and activations are omitted for brevity.

general signal for paraphrase representation learning. COMET (Rei et al., 2020) is a framework for training multilingual machine translation (MT) evaluation models where parameters in the regression or ranking layers are optimized using human judgements scores with either an MSE objective or triplet objective respectively. PRISM (Thompson and Post, 2020) similar to ParaBLEU formulates evaluation as a paraphrasing task. However it treats paraphrasing as zero-shot translation using a multilingual neural MT model as a paraphraser. BARTScore (Yuan et al., 2021) calculates the log-likelihood of the candidate text conditioned upon the reference text from BART (Lewis et al., 2019), a pretrained sequence-to-sequence model.

Paraphrase generation There is a wealth of recent literature on controllable paraphrase generation and linguistic style transfer (Yang et al., 2018; Zhao et al., 2018; Jin et al., 2020), which aims to extract the style of a piece of text and map it onto another piece of text without changing its semantic meaning. T5 leverages a huge text corpus as pretraining for conditional generation using 'commands' encoded as text, which includes paraphrastic tasks such as summarization. FSET (Kazemnejad et al., 2020) is a retrieval-based paraphrase generation system in which a sentence z is paraphrased by first locating a similar reference sentence from a large bank of reference/candidate pairs, then extracting and replaying similar low-level edits on z. Common to ParaBLEU and FSET is the use of a Transformer for paraphrase style transfer, with differing architectural details. However, FSET is designed to transpose low-level edits and so requires lexically similar examples; whereas ParaBLEU is explicitly designed to learn high-level, reference-invariant paraphrase representations using a factorized objective. The musical style Transformer autoencoder (Choi et al., 2020) uses a similar Transformer-based style transfer architecture to conditionally generate new music in controllable styles. Other examples in text-to-speech systems perform style transfer by encoding the prosody of a source sentence into a bottlenecked reference embedding (Skerry-Ryan et al., 2018) or disentangled style tokens (Wang et al., 2018b). STRAP (Krishna et al., 2020) generates paraphrases in controllable styles by mixing and matching multiple style-specific fine-tuned GPT-2 models. REAP (Goyal and Durrett, 2020) uses a Transformer to generate syntactically diverse paraphrases by including an additional position embedding representing the syntactic tree. DNPG (Li et al., 2019) is a paraphrase generation system that uses a cascade of Transformer encoders/decoders to control whether paraphrasing is sentential/phrasal.

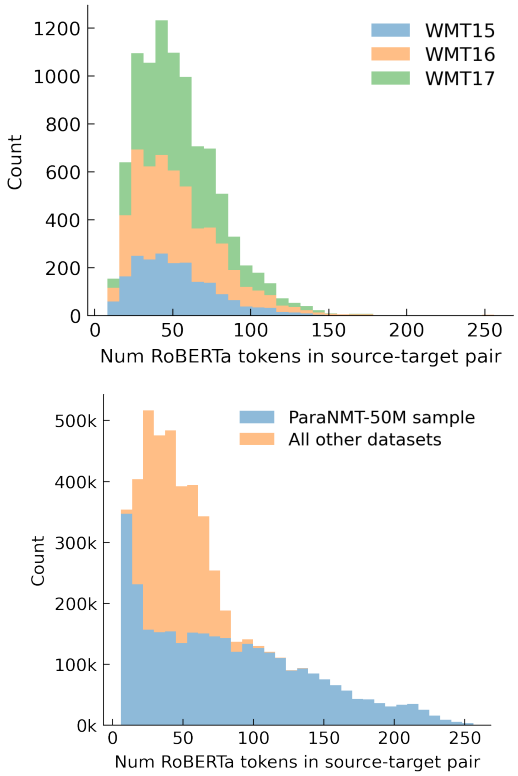

Figure 2: Stacked histograms showing the distribution of the number of RoBERTa tokens in the WMT Metrics Shared Task data (top) and ParaCorpus (bottom).

3 Data

In this section, we describe the pretraining and fine-tuning datasets we use in our studies.

3.1 WMT Metrics Shared Task

The WMT Metrics Shared Task is an annual benchmark for automated evaluation metrics for translation systems, where the goal is to predict average human ratings comparing the machine-translated candidate $\hat{x}$ with human-translated reference x, both of which have been translated from the same source sentence.

We use an identical setup to (Sellam et al., 2020) and (Zhang et al., 2019a), where we use the subset of data for which the candidate and reference are in English, which we will refer to as the to-English subset. The source, which is unused, can be in any non-English language, the set of which varies from year-to-year. We produce results for the WMT Metrics Shared Task 2017 (WMT17), training on the to-English subsets of WMT15 and WMT16. The test set contains $4,132$ examples and the training set $5,360$ examples. The distributions of example length in tokens is shown in Figure 2. The WMT data is prepared using the WMT preparation code in the BLEURT repository[1]. The decision to test only on the WMT17 dataset is deliberate. Results from previous state-of-the-art papers (Sellam et al., 2020; Zhang et al., 2019a) demonstrate issues with WMT18 and later datasets: the noise in the test set is high and differentiation between different methods becomes so suppressed for later years that the benchmark becomes uninteresting. This issue is noted in both the BLEURT paper and by the organizers of the 2018 WMT Metrics Shared Task[2].

We report the agreement between the metric and the human scores using two related correlation coefficients: absolute Kendall $|\tau|$ and absolute Pearson $|r|$, the latter of which was the official metric of the 2017 task. In our summary results in the main paper, we average these metrics across all source languages but not over reference/candidate language. Full results are provided in Appendix E.

3.2 ParaCorpus

In addition to our design choices, we also encourage a robust and generalizable pretraining by using a dataset that covers a variety of styles and lengths. We collate a number of paraphrase datasets to create a single pretraining dataset we call ParaCorpus. The composition of the dataset is shown in Table 4, with a total of $\sim$ 5.1m examples. All examples have reference and candidate texts and around one third additionally have binary entailment labels. Where the source dataset included three-way labels 'entailment'/'contradiction'/'neutral', 'entailment' was mapped to 1 and the others to 0. A subset of ParaNMT-50M (Wieting and Gimpel, 2017), which includes noisier, speech-like examples, was included to add additional stylistic diversity to the dataset, and to increase the population of the dataset with combined token lengths above 128, which we hypothesize will make the model more robust to the longer examples seen in the WMT datasets. Token lengths are shown in Figure 2.

4 Experiments

In this section, we present results on WMT17, benchmarked against the current state-of-the-art approach, along with widely used neural, n-gram and edit-distance-based metrics. We study ParaBLEU performance as a function of number of pretraining

[1] https://github.com/google-research/bleurt

[2] https://www.statmt.org/wmt18/metrics-task.html

Table 1: Summary results for WMT17. The metrics reported are absolute Kendall $|\tau|$ and Pearson $|r|$ averaged across each source language. Full results can be found in Appendix E.

Model	$\lvert\tau\rvert$	$\lvert r\rvert$
BLEU	0.292	0.423
TER	0.352	0.475
ROUGE	0.354	0.518
METEOR	0.301	0.443
chrF++	0.396	0.578
BLEURT-large	0.625	0.818
BERTScore-RoBERTa$_{\text{large}}$	0.567	0.759
BERTScore-T5$_{\text{large}}$	0.536	0.738
BERTScore-DeBERTa$_{\text{large}}$	0.580	0.773
MoverScore	0.322	0.454
ParaBLEU$_{\text{large}}$	**0.653**	**0.843**
ParaBLEU$_{\text{base}}$	0.589	0.785

steps and the size of the fine-tuning dataset. Finally, we perform ablations to test the impact of the inductive biases and resultant architectural decisions described in Section 2.

We report results for both ParaBLEU$_{\text{base}}$, based on RoBERTa$_{\text{base}}$ (12 layers, 768 hidden units, 12 heads), and our default model ParaBLEU$_{\text{large}}$, based on RoBERTa$_{\text{large}}$ (24 layers, 1,024 hidden units, 16 heads). Both models are trained near-identically for 4 epochs on ParaCorpus. Further pretraining details can be found in Appendix A. For fine-tuning, we use a batch size of 32, a learning rate of 1e-5 and train for 40k steps, with a validation set size of 10% (unless otherwise stated). No reference texts are shared between the train and validation sets, following (Sellam et al., 2020). Pretraining ParaBLEU$_{\text{large}}$ takes $\sim$ 10h on a 16 A100 GPU machine. Fine-tuning takes $\sim$ 8h on a single A100 GPU machine.

4.1 Results

ParaBLEU results on WMT17 are given in Table 1, along with a number of baselines described in Section 2.3).

ParaBLEU$_{\text{large}}$ achieves new state-of-the-art results on WMT17, exceeding the previous state-of-the-art approach, BLEURT, on both correlation metrics. We note that non-neural metrics perform the worst, of which the character-level n-gram-matching algorithm chrF++ performs the best. Non-learned neural metrics (BERTScore and MoverScore) tend to perform better, and learned neural metrics (BLEURT and ParaBLEU) perform the best. BLEU, the most widely used metric, has the poorest correlation with human judgements. This is consistent with results seen previously in the literature (Zhang et al., 2019a; Sellam et al., 2020). The significant drop in performance from ParaBLEU$_{\text{large}}$ to ParaBLEU$_{\text{base}}$ highlights the benefit of larger, more expressive pretrained language models.

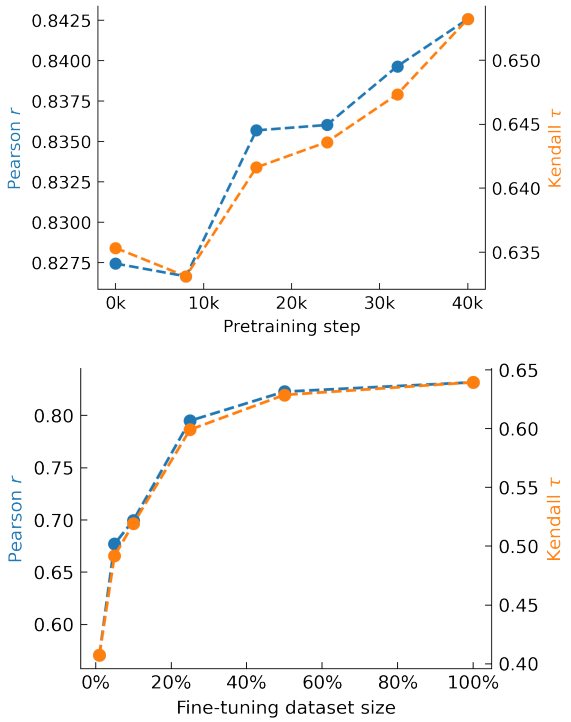

Figure 3: Performance of ParaBLEU$_{\text{large}}$ on WMT17 as a function of number of pretraining steps (top) and the fine-tuning dataset size (bottom). Note that the Pearson r results (blue) use the left y-axis, whereas Kendall τ (orange) uses the right y-axis.

Figure 3 probes performance as a function of number of pretraining steps and the size of the fine-tuning dataset for ParaBLEU$_{\text{large}}$. As expected, pretraining for longer increases downstream task performance. However, we note that 40k steps, approximately 4 epochs of ParaCorpus, does not yet reach diminishing returns on WMT17 performance. We therefore recommend pretraining for significantly longer. Both BERT and RoBERTa are pretrained for 40 epochs (Liu et al., 2019; Lan et al., 2019); the T5 authors ablate their dataset size at a fixed number of steps and conclude that performance does not significantly degrade up to and including 64 epochs (Raffel et al., 2019); conversely,

Table 2: Ablation results on WMT17. The metrics reported are the absolute Kendall $|\tau|$ and Pearson $|r|$ correlation coefficients averaged across each reference language.

Model	$\lvert\tau\rvert$	$\lvert r\rvert$
Baseline (ParaBLEU$_{\text{large}}$)	**0.653**	**0.843**
No MLM loss ($\mathcal{L}_{\text{MLM}}$)	0.633	0.826
No autoregressive loss ($\mathcal{L}_{\text{AR}}$)	0.642	0.834
No entailment classification loss ($\mathcal{L}_{\text{CLS}}$)	0.644	0.837

the BLEURT authors see diminishing returns on downstream task performance after 2 pretraining epochs (Sellam et al., 2020).

For the fine-tuning dataset size study, we consistently use a validation set size of 25% to facilitate the small-data results. Despite the training set (the to-English subsets of WMT15 and WMT16) forming a relatively small dataset, ParaBLEU$_{\text{large}}$ trained on 50% of the available data (2,010 training examples, 670 validation examples) still beats the previous state-of-the-art, BLEURT, yielding a Pearson correlation of 0.823. The impact of reducing the train size from 100% (4,020 training examples, 1,340 validation examples) to 25% (1,005 training examples, 335 validation examples) has a relatively small effect on performance, reducing Pearson r from 0.832 to 0.795. With a dataset size of only 1% (40 training examples, 14 validation examples), ParaBLEU$_{\text{large}}$ achieves a Pearson r of 0.571, still correlating significantly more strongly with human judgements than BLEU, TER, ROUGE, METEOR and MoverScore. We attribute this to the suitability of the generalized pretraining objective for priming the model for paraphrase evaluation tasks.

4.2 Ablations

To more directly test the hypotheses in Section 2.1, we perform ablations in which we remove each component of the factorized objective in turn. The results of this are shown in Table 2. Each part of the objective is associated with an increase in downstream task performance. The most significant degradation comes from removing the MLM loss. Possible reasons for this include: the MLM loss' contribution to candidate acceptability judgement are crucial; the MLM loss acts as a regularizer, encouraging the edit encoder to represent paraphrases in linguistic concepts rather than low-level edits; and the MLM loss further encourages bitext alignment behaviour, as described in Section 2.1.

5 One-shot paraphrase generation

As our final study, we exploit the generative nature of the pretraining architecture to test our claim that the edit encoder reasons in high-level paraphrastic concepts rather than low-level edits. To do this, we diverge from the pretraining setup, in which the same reference text is passed to both the edit encoder and the sequence-to-sequence model, by passing a different, unseen reference to the sequence-to-sequence model. Akin to (Brown et al., 2020; Gao et al., 2020), the hope is that the 'demonstration paraphrase' acts as a conditioning factor for paraphrasing the unseen sentence in a similar way.

If the model is reasoning in low-level edits or otherwise 'cheating', we expect to see:

- Thematic/word leakage from the encoder candidate to the generated candidate, caused by the candidate being autoencoded. This is the undesirable behaviour we sought to address using a bottleneck.
- Ungrammatical or otherwise unacceptable output with made-up words and/or bad word order, caused by the encoding of low-level edits scrambling the generator reference tokens.

If the model is reasoning in high-level paraphrastic concepts, we expect to see:

- Consistently grammatical, acceptable output.
- The flavour of the paraphrase mirroring the conditioning, e.g. the altering of a linguistic style, mood or tense.

We generate text using beam-search (Medress et al., 1977). We sample references at random from the MRPC dataset. The demonstration candidate is a hand-crafted paraphrase of the demonstration reference that embodies a pre-specified paraphrase type. We report the predicted entailment score of the demonstration reference and candidate, along with the candidate generated by the model.

A summarized, random subset of generation results is shown in Appendix C. We include two sets of results for each paraphrastic type (e.g.

'negative'): one where the demonstration reference/candidate differ in this concept, and one where both embody the concept. Since we wish to encode the *difference* between the demonstration reference/candidate texts, the desired behaviour when the demonstration pair is identical is no change. If this is not the case, it is likely that the edit encoder is just autoencoding the candidate using high-level linguistic concepts, similar to linguistic style transfer. Further randomly chosen examples are given in Appendix F.

The results present a strong case that the encoder is representing high-level paraphrastic concepts. It is able to successfully identify changes in mood, style and tense between the demonstration reference and candidate, and transpose them onto the unseen reference to make a largely grammatical and appropriately paraphrased sentence. We do not see significant leakage of concepts, words or styles between the demonstration candidate and the generated candidate, instead the expected transfer of paraphrase style.

6 Limitations

Limitations of this work include the relatively small set of baselines used; there is an ever-increasing number of text similarity metrics and so only a subset is presented here. As demonstrated in Section 4.1, it seems likely that the performance is currently limited by pretraining time and so we have not yet probed the ceiling performance of this method. Swapping out the edit encoder and encoder-decoder for current state-of-the-art models like DeBERTa (He et al., 2020) may offer further performance boosts. Expanding this work to predicting on datasets beyond the 2017 WMT Metrics Shared Task will probe the generalizability of the techniques in this paper. The application of ParaBLEU for paraphrase generation has not been quantitatively explored. Although the results presented in Section 5 are chosen at random, the analysis is qualitative. More rigorous methods for evaluating the quality of the paraphrase generation are left for future work.

7 Conclusions

In this paper, we introduced ParaBLEU, a paraphrase representation learning model and associated paraphrase evaluation metric. We demonstrated that the metric yields state-of-the-art correlation with human paraphrase judgements and is robust to data scarcity. We motivated its pretraining strategy through a set of inductive biases, which we tested through ablation studies. Finally, we reframed the pretraining as a one-shot paraphrase generation model and gathered evidence that ParaBLEU represents meaningful paraphrastic information.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Kristy Choi, Curtis Hawthorne, Ian Simon, Monica Dinculescu, and Jesse Engel. 2020. Encoding musical style with transformer autoencoders. In *International Conference on Machine Learning*, pages 1899–1908. PMLR.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. Parasci: A large scientific paraphrase dataset for

longer paraphrase generation. *arXiv preprint arXiv:2101.08382*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Tanya Goyal and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. *arXiv preprint arXiv:2005.02013*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2020. Deep learning for text style transfer: A survey. *CoRR, abs/2011.00416*.

David K Johnson, Martha Storandt, and David A Balota. 2003. Discourse analysis of logical memory recall in normal aging and in dementia of the alzheimer type. *Neuropsychology*, 17(1):82.

Amirhossein Kazemnejad, Mohammadreza Salehi, and Mahdieh Soleymani Baghshah. 2020. Paraphrase generation by learning how to edit from samples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6010–6021.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. *arXiv preprint arXiv:2010.05700*.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, pages 707–710. Soviet Union.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Zichao Li, Xin Jiang, Lifeng Shang, and Qun Liu. 2019. Decomposable neural paraphrase generation. *arXiv preprint arXiv:1906.09741*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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*.

Mark F. Medress, Franklin S Cooper, Jim W. Forgie, CC Green, Dennis H. Klatt, Michael H. O'Malley, Edward P Neuburg, Allen Newell, DR Reddy, B Ritea, et al. 1977. Speech understanding systems: Report of a steering committee. *Artificial Intelligence*, 9(3):307–316.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

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*, pages 160–167.

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.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

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

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with

tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. *arXiv preprint arXiv:2004.14564*.

Jörg Tiedemann. 2011. Bitext alignment. *Synthesis Lectures on Human Language Technologies*, 4(2):1–165.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018a. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018b. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pages 5180–5189. PMLR.

Sandra Weintraub, Lilah Besser, Hiroko H Dodge, Merilee Teylan, Steven Ferris, Felicia C Goldstein, Bruno Giordani, Joel Kramer, David Loewenstein, Dan Marson, et al. 2018. Version 3 of the alzheimer disease centers' neuropsychological test battery in the uniform data set (uds). *Alzheimer disease and associated disorders*, 32(1):10.

John Wieting and Kevin Gimpel. 2017. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. *arXiv preprint arXiv:1805.11749*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *arXiv preprint arXiv:2106.11520*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

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. *arXiv preprint arXiv:1909.02622*.

Yanpeng Zhao, Wei Bi, Deng Cai, Xiaojiang Liu, Kewei Tu, and Shuming Shi. 2018. Language style transfer from sentences with arbitrary unknown styles. *arXiv preprint arXiv:1808.04071*.

A Pretraining hyperparameters

Table 3 shows the hyperparamaters used for the $\text{ParaBLEU}_{\text{base}}$ and $\text{ParaBLEU}_{\text{large}}$ models during pretraining. α and β are the loss weights from Equation 1.

B ParaCorpus

Table 4 provides a description of the composition of the pretraining dataset.

C One-shot paraphrase generation results

Table 5 presents the results from our one-shot paraphrase generation experiment detailed in Section 5.

D Microsoft Research Paraphrase Corpus results

We additionally ran a study on the Microsoft Research Paraphrase Corpus (MRPC) (Dolan and Brockett, 2005), a constituent of the GLUE benchmark (Wang et al., 2018a). MRPC contains $5,801$ sentence pairs each accompanied by hand-labelled binary judgement of whether the pair constitutes a paraphrase. The data is split into a train set ($4,076$ sentence pairs of which $2,753$ are paraphrases) and a test set ($1,725$ sentence pairs of which $1,147$ are paraphrases).

Hyperparameter	ParaBLEU$_{\text{base}}$	ParaBLEU$_{\text{large}}$
Edit encoder base model	RoBERTa$_{\text{base}}$	RoBERTa$_{\text{large}}$
Sequence-to-sequence base model	BART$_{\text{base}}$	BART$_{\text{base}}$
Batch size (per GPU; examples)	64	32
Batch size (per GPU; max tokens)	16,384	8,192
Learning rate (per GPU)	4e-4	1e-4
Warmup steps	1,200	2,400
Train length (updates)	20k	40k
Train length (epochs)	4	4
Gradient accumulation steps	1	2
α	2.0	2.0
β	10.0	10.0

Table 3: Pretraining hyperparameters for the ParaBLEU$_{\text{base}}$ and ParaBLEU$_{\text{large}}$ models used in this paper. These were adapted for a larger architecture from the RoBERTa paper (Liu et al., 2019) and not subject to tuning.

We fine-tune our ParaBLEU models on the MRPC train set using the fine-tuning procedure detailed in (Liu et al., 2019) and predict on the held-out test set. For baselines we use the ALBERT$_{\text{large}}$ (Lan et al., 2019) and the RoBERTa$_{\text{large}}$ (Liu et al., 2019) models fine-tuned using their respective hyperparameters.

Model	Accuracy	F1 score
ALBERT$_{\text{large}}$	88.2	91.3
RoBERTa$_{\text{large}}$	89.5	92.2
ParaBLEU$_{\text{large}}$	88.8	91.5
ParaBLEU$_{\text{base}}$	85.2	88.9

Table 6: The results from the Microsoft Research Paraphrase Corpus (MRPC).

From Table 6 we observe that our default model ParaBLEU$_{\text{large}}$ underperforms compared to the model it is based on, RoBERTa$_{\text{large}}$. This could be because the hyperparameter sweep we used for our ParaBLEU models (the same sweep as recommended by the authors of RoBERTa$_{\text{large}}$) is suboptimal and a broader hyperparameter sweep may be required.

Table 4: ParaCorpus composition. † indicates that the dataset does not have a binary 'entailment' label, but instead has a three-way entail/contradict/neutral label which we map to a binary 'entailment' label as described in Section 3.2.

Dataset	Subsets included	Nature	Size	Ent. labels	Ref
PAWS	Wiki-train; QQP-train	Sentence pairs with high semantic overlap	740k	✓	(Zhang et al., 2019b)
SNLI	Train	Human-written entailment sentence pairs	550k	✓†	(Bowman et al., 2015)
MultiNLI	Train	Multi-genre entailment sentence pairs	390k	✓†	(Williams et al., 2017)
ParaSCI	ACL-train; arXiv-train	Human-written academic paraphrase pairs	340k	✗	(Dong et al., 2021)
ParaNMT-50M	Random sample (see main text)	Varied paraphrase pairs from machine translation	3.1m	✗	(Wieting and Gimpel, 2017)
ParaCorpus	-	-	5.1m	Partial	-

Table 5: One-shot paraphrase generation results. In every case, the reference is *"Congo's war began in 1998 when Uganda and Rwanda invaded to back rebels fighting to topple the central government."* The difference between the identity and non-identity version is shown in boldface. 'Ent.' is the predicted demonstration reference/candidate entailment probability from the classification loss.

Type	Demo reference	Demo candidate	Ent.	Generated candidate
negation	Both bidders agreed to assume about $90 million in debt owed on the planes	The bidders did not agree to assume about $90 million in debt owed on the planes	0.0%	The Congo War began in 1998 when Uganda and Rwanda **did not invade** to back rebels fighting to topple the central government.
identity (negation)	The bidders did not agree to assume about $90 million in debt owed on the planes	The bidders did not agree to assume about $90 million in debt owed on the planes	92.6%	The Congo War began in 1998 when Uganda and Rwanda invaded to back rebels fighting to topple the central government.
summary	Both bidders agreed to assume about $90 million in debt owed on the planes	They agreed to assume the debt	99.0%	**They invaded to back** the rebels fighting to topple the central government.
identity (summary)	They agreed to assume the debt	They agreed to assume the debt	95.6%	Congo began in 1998 when Uganda and Rwanda invaded to support the rebels fighting to topple the central government.
passive	Both bidders agreed to assume about $90 million in debt owed on the planes	$90 million in debt owed on the planes was agreed to be assumed by both bidders	99.8%	**The war in Congo** began in 1998 when Uganda and Rwanda **were invaded to back** rebels fighting to topple the central government.
identity (passive)	$90 million in debt owed on the planes was agreed to be assumed by both bidders	$90 million in debt owed on the planes was agreed to be assumed by both bidders	92.7%	Congo War began in 1998 when Uganda and Rwanda invaded to back rebels fighting to topple the central government.
conditional perfect	Both bidders agreed to assume about $90 million in debt owed on the planes	Both bidders would have agreed to assume about $90 million in debt owed on the planes	69.4%	**Congo's war** began in 1998, when Uganda and Rwanda **would have** invaded to back rebels fighting to topple the central government.
identity (conditional perfect)	Both bidders would have agreed to assume about $90 million in debt owed on the planes	Both bidders would have agreed to assume about $90 million in debt owed on the planes	90.7%	Congo War began in 1998 when Uganda and Rwanda invaded to back rebels fighting to topple the central government.
personal	Both bidders agreed to assume about $90 million in debt owed on the planes	We agreed to assume about $90 million in debt owed on the planes	95.4%	**We started the Congo war** in 1998 when Uganda and Rwanda invaded to support rebels fighting to topple the central government.
identity (personal)	We agreed to assume about $90 million in debt owed on the planes	We agreed to assume about $90 million in debt owed on the planes	93.2%	The Congo War began in 1998 when Uganda and Rwanda invaded to back rebels fighting to topple the central government.
informal	Both bidders agreed to assume about $90 million in debt owed on the planes	Both guys agreed to take on like $90 million that they owed for some planes	98.8%	**Congo's war started** in 1998 when Uganda and Rwanda invaded to **get some** rebels fighting to topple the central government.
identity (informal)	Both guys agreed to take on like $90 million that they owed for some planes	Both guys agreed to take on like $90 million that they owed for some planes	94.5%	Congo War began in 1998 when Uganda and Rwanda invaded to back rebels fighting to topple the central government.

E Full to-English WMT results

Table 7 shows the full WMT17 results, which are summarized in main paper Table 1. See Section 4.1 for more details.

Table 7: Full to-English results for WMT17. The metrics reported are absolute Kendall $|\tau|$ and Pearson $|r|$. Models are fine-tuned on the English subset of WMT15 and WMT16. For a language pair 'x-y', the original reference was in language 'x', and both human and machine translations are in language 'y'. For results averaged across all source languages, see the main paper.

Model	lv-en $\lvert\tau\rvert$ / $\lvert r\rvert$	tr-en $\lvert\tau\rvert$ / $\lvert r\rvert$	zh-en $\lvert\tau\rvert$ / $\lvert r\rvert$	ru-en $\lvert\tau\rvert$ / $\lvert r\rvert$	de-en $\lvert\tau\rvert$ / $\lvert r\rvert$	cs-en $\lvert\tau\rvert$ / $\lvert r\rvert$	fi-en $\lvert\tau\rvert$ / $\lvert r\rvert$
BLEU	0.215 / 0.334	0.313 / 0.461	0.344 / 0.488	0.313 / 0.431	0.259 / 0.372	0.255 / 0.373	0.342 / 0.503
TER	0.329 / 0.439	0.393 / 0.472	0.365 / 0.493	0.358 / 0.509	0.295 / 0.403	0.315 / 0.458	0.411 / 0.548
ROUGE	0.303 / 0.459	0.395 / 0.56	0.366 / 0.542	0.343 / 0.488	0.336 / 0.488	0.302 / 0.462	0.434 / 0.628
METEOR	0.258 / 0.403	0.375 / 0.554	0.352 / 0.521	0.353 / 0.491	0.307 / 0.445	0.287 / 0.448	0.402 / 0.597
MoverScore	0.252 / 0.350	0.314 / 0.493	0.345 / 0.485	0.375 / 0.493	0.296 / 0.401	0.317 / 0.433	0.356 / 0.521
chrF++	0.333 / 0.520	0.432 / 0.614	0.405 / 0.593	0.415 / 0.588	0.365 / 0.534	0.35 / 0.523	0.475 / 0.678
BLEURT	0.644 / 0.835	0.629 / 0.824	0.602 / 0.814	0.613 / 0.811	0.599 / 0.792	0.593 / 0.773	0.695 / 0.878
BERTScore-RoBERTa$_{\text{large}}$	0.555 / 0.756	0.569 / 0.751	0.568 / 0.775	0.555 / 0.746	0.554 / 0.745	0.522 / 0.71	0.646 / 0.833
BERTScore-T5$_{\text{large}}$	0.529 / 0.74	0.53 / 0.721	0.532 / 0.749	0.531 / 0.74	0.5 / 0.699	0.485 / 0.69	0.643 / 0.831
BERTScore-DeBERTa$_{\text{large}}$	0.581 / 0.785	0.579 / 0.755	0.584 / 0.795	0.576 / 0.771	0.561 / 0.751	0.537 / 0.729	0.642 / 0.825
ParaBLEU$_{\text{large}}$	0.641 / 0.832	0.643 / 0.846	0.586 / 0.791	0.628 / 0.824	0.612 / 0.796	0.607 / 0.797	0.695 / 0.881
ParaBLEU$_{\text{base}}$	0.603 / 0.805	0.627 / 0.824	0.565 / 0.777	0.580 / 0.780	0.568 / 0.764	0.530 / 0.704	0.649 / 0.838

F More generation examples

This section includes additional examples of one-shot paraphrase generation sampled from the MRPC dataset. See Section 5 for more information.

Table 8: Additional one-shot paraphrase generation results. 'Ent. score' is the predicted demonstration reference/candidate entailment probability from the classification loss. See Section 5 for more information.

#	Type	Demo reference	Demo candidate	Ent. score	Reference	Generated candidate
1	identity	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both bidders agreed to assume about $ 90 million in debt owed on the planes	90.0%	Robert Stewart , a spokesman for Park Place , the parent company of Caesars Palace , said he was surprised by the court 's decision .	Robert Stewart, a spokesman for Park Place, the parent company of Caesars Palace, said he was surprised by the court decision.
2	identity	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both bidders agreed to assume about $ 90 million in debt owed on the planes	90.0%	Democrats now hope to increase the value of awards proposed by Hatch and to create a mechanism to ensure the fund remains solvent .	Democrats now hope to increase the value of awards proposed by Hatch and create a mechanism to ensure that the fund remains solvent.
3	identity	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both bidders agreed to assume about $ 90 million in debt owed on the planes	90.0%	Indonesia 's army has often been accused of human rights abuses during GAM 's battle for independence , charges it has generally denied while accusing the separatists of committing rights violations .	Indonesia's army has often been accused of human rights abuses during GAM's battle for independence, charges it generally denied while accusing the separatists of committing rights violations.
4	identity	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both bidders agreed to assume about $ 90 million in debt owed on the planes	90.0%	Progress Software plans to acquire privately held DataDirect Technologies for about $ 88 million in cash , the companies said Friday .	Progress Software plans to acquire privately held DataDirect Technologies for about $ 88 million in cash, the companies said Friday.
5	identity	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both bidders agreed to assume about $ 90 million in debt owed on the planes	90.0%	A positive PSA test has to be followed up with a biopsy or other procedures before cancer can be confirmed .	A positive PSA test must be followed up with biopsy or other procedures before cancer can be confirmed.

Table 9: Additional one-shot paraphrase generation results. 'Ent. score' is the predicted demonstration reference/candidate entailment probability from the classification loss.

#	Type	Demo reference	Demo candidate	Ent. score	Reference	Generated candidate
6	negation	Both bidders agreed to assume about $ 90 million in debt owed on the planes	The bidders did not agree to assume about $ 90 million in debt owed on the planes	0.0%	Robert Stewart , a spokesman for Park Place , the parent company of Caesars Palace , said he was surprised by the court 's decision .	Robert Stewart, a spokesman for Park Place, the parent company of Caesars Palace, did not say he was surprised by the court decision.
7	negation	Both bidders agreed to assume about $ 90 million in debt owed on the planes	The bidders did not agree to assume about $ 90 million in debt owed on the planes	0.0%	Democrats now hope to increase the value of awards proposed by Hatch and to create a mechanism to ensure the fund remains solvent .	Democrats now hope to increase the value of awards proposed by Hatch and to create a mechanism to ensure the fund does not remain solvent.
8	negation	Both bidders agreed to assume about $ 90 million in debt owed on the planes	The bidders did not agree to assume about $ 90 million in debt owed on the planes	0.0%	Indonesia 's army has often been accused of human rights abuses during GAM 's battle for independence , charges it has generally denied while accusing the separatists of committing rights violations .	The Indonesian army did not often be accused of human rights abuses during the GAM struggle for independence, charges it generally denied while accusing the separatists of committing rights violations.
9	negation	Both bidders agreed to assume about $ 90 million in debt owed on the planes	The bidders did not agree to assume about $ 90 million in debt owed on the planes	0.0%	Progress Software plans to acquire privately held DataDirect Technologies for about $ 88 million in cash , the companies said Friday .	Progress Software plans to acquire the privately held DataDirect Technologies for about $ 88 million in cash, the companies said Friday.
10	negation	Both bidders agreed to assume about $ 90 million in debt owed on the planes	The bidders did not agree to assume about $ 90 million in debt owed on the planes	0.0%	A positive PSA test has to be followed up with a biopsy or other procedures before cancer can be confirmed .	A positive PSA test does not have to be followed up with biopsy or other procedures before cancer can be confirmed.

Table 10: Additional one-shot paraphrase generation results. 'Ent. score' is the predicted demonstration reference/candidate entailment probability from the classification loss.

#	Type	Demo reference	Demo candidate	Ent. score	Reference	Generated candidate
11	summary	Both bidders agreed to assume about $ 90 million in debt owed on the planes	They agreed to assume the debt	99.0%	Robert Stewart , a spokesman for Park Place , the parent company of Caesars Palace , said he was surprised by the court 's decision .	They said they were surprised by the decision.
12	summary	Both bidders agreed to assume about $ 90 million in debt owed on the planes	They agreed to assume the debt	99.0%	Democrats now hope to increase the value of awards proposed by Hatch and to create a mechanism to ensure the fund remains solvent .	They hoped to increase the value of the awards.
13	summary	Both bidders agreed to assume about $ 90 million in debt owed on the planes	They agreed to assume the debt	99.0%	Indonesia 's army has often been accused of human rights abuses during GAM 's battle for independence , charges it has generally denied while accusing the separatists of committing rights violations .	They generally accused the separatists of committing rights violations.
14	summary	Both bidders agreed to assume about $ 90 million in debt owed on the planes	They agreed to assume the debt	99.0%	Progress Software plans to acquire privately held DataDirect Technologies for about $ 88 million in cash , the companies said Friday .	They wanted to acquire DataDirect Technologies.
15	summary	Both bidders agreed to assume about $ 90 million in debt owed on the planes	They agreed to assume the debt	99.0%	A positive PSA test has to be followed up with a biopsy or other procedures before cancer can be confirmed .	They had to confirm the PSA test.

Table 11: Additional one-shot paraphrase generation results. 'Ent. score' is the predicted demonstration reference/candidate entailment probability from the classification loss.

#	Type	Demo reference	Demo candidate	Ent. score	Reference	Generated candidate
16	passive	Both bidders agreed to assume about $ 90 million in debt owed on the planes	$ 90 million in debt owed on the planes was agreed to be assumed by both bidders	99.8%	Robert Stewart , a spokesman for Park Place , the parent company of Caesars Palace , said he was surprised by the court 's decision .	The court decision was surprised by Robert Stewart, a spokesman for Park Place, the parent company of Caesars Palace.
17	passive	Both bidders agreed to assume about $ 90 million in debt owed on the planes	$ 90 million in debt owed on the planes was agreed to be assumed by both bidders	99.8%	Democrats now hope to increase the value of awards proposed by Hatch and to create a mechanism to ensure the fund remains solvent .	The value of awards proposed by Hatch is now hoped to increase by Democrats and create a mechanism to ensure the fund remains solvent.
18	passive	Both bidders agreed to assume about $ 90 million in debt owed on the planes	$ 90 million in debt owed on the planes was agreed to be assumed by both bidders	99.8%	Indonesia 's army has often been accused of human rights abuses during GAM 's battle for independence , charges it has generally denied while accusing the separatists of committing rights violations .	Human rights abuses in the battle for independence was often accused by Indonesia's army, charges which it generally denied while accusing the separatists of committing rights violations.
19	passive	Both bidders agreed to assume about $ 90 million in debt owed on the planes	$ 90 million in debt owed on the planes was agreed to be assumed by both bidders	99.8%	Progress Software plans to acquire privately held DataDirect Technologies for about $ 88 million in cash , the companies said Friday .	The privately held DataDirect Technologies plans to be acquired by Progress Software in cash for about $ 88 million.
20	passive	Both bidders agreed to assume about $ 90 million in debt owed on the planes	$ 90 million in debt owed on the planes was agreed to be assumed by both bidders	99.8%	A positive PSA test has to be followed up with a biopsy or other procedures before cancer can be confirmed .	A positive PSA test has to be confirmed in biopsy or other procedures before cancer can be followed up.

Table 12: Additional one-shot paraphrase generation results. 'Ent. score' is the predicted demonstration reference/candidate entailment probability from the classification loss.

#	Type	Demo reference	Demo candidate	Ent. score	Reference	Generated candidate
21	conditional perfect	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both bidders would have agreed to assume about $ 90 million in debt owed on the planes	69.4%	Robert Stewart , a spokesman for Park Place , the parent company of Caesars Palace , said he was surprised by the court 's decision .	Robert Stewart, a spokesman for Caesars Palace, the parent company of Park Place, said he would have been surprised by the court decision.
22	conditional perfect	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both bidders would have agreed to assume about $ 90 million in debt owed on the planes	69.4%	Democrats now hope to increase the value of awards proposed by Hatch and to create a mechanism to ensure the fund remains solvent .	Democrats would now hope to increase the value of awards proposed by Hatch and to create a mechanism to ensure that the fund would remain solvent.
23	conditional perfect	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both bidders would have agreed to assume about $ 90 million in debt owed on the planes	69.4%	Indonesia 's army has often been accused of human rights abuses during GAM 's battle for independence , charges it has generally denied while accusing the separatists of committing rights violations .	Indonesia's army would have often been accused of human rights abuses during GAM's battle for independence, charges which it generally denied while accusing the separatists of committing rights violations.
24	conditional perfect	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both bidders would have agreed to assume about $ 90 million in debt owed on the planes	69.4%	Progress Software plans to acquire privately held DataDirect Technologies for about $ 88 million in cash , the companies said Friday .	Progress Software plans to acquire privately held DataDirect Technologies for about $ 88 million in cash, the companies said Friday.
25	conditional perfect	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both bidders would have agreed to assume about $ 90 million in debt owed on the planes	69.4%	A positive PSA test has to be followed up with a biopsy or other procedures before cancer can be confirmed .	A positive PSA test would have to be followed up with a biopsy or other procedures before cancer can be confirmed.

Table 13: Additional one-shot paraphrase generation results. 'Ent. score' is the predicted demonstration reference/candidate entailment probability from the classification loss.

#	Type	Demo reference	Demo candidate	Ent. score	Reference	Generated candidate
26	unrelated	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Bush turned out a statement yesterday thanking the commission for its work , and said , " Our journey into space will go on . "	0.2%	Robert Stewart , a spokesman for Park Place , the parent company of Caesars Palace , said he was surprised by the court 's decision .	A spokesman for Caesars Palace, the parent company of Park Place, said he was surprised by the court's decision.
27	unrelated	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Bush turned out a statement yesterday thanking the commission for its work , and said , " Our journey into space will go on . "	0.2%	Democrats now hope to increase the value of awards proposed by Hatch and to create a mechanism to ensure the fund remains solvent .	Hatch now hopes to raise the value of awards proposed by Democrats, and create a mechanism to ensure that the fund remains solvent.
28	unrelated	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Bush turned out a statement yesterday thanking the commission for its work , and said , " Our journey into space will go on . "	0.2%	Indonesia 's army has often been accused of human rights abuses during GAM 's battle for independence , charges it has generally denied while accusing the separatists of committing rights violations .	During the battle for independence, GAM's army has often been accused of human rights abuses, which it generally denied, while accusing the separatists of committing rights violations.
29	unrelated	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Bush turned out a statement yesterday thanking the commission for its work , and said , " Our journey into space will go on . "	0.2%	Progress Software plans to acquire privately held DataDirect Technologies for about $ 88 million in cash , the companies said Friday .	The company announced on Friday that Progress Software plans to acquire DataDirect Technologies, for about $ 88 million in cash.
30	unrelated	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Bush turned out a statement yesterday thanking the commission for its work , and said , " Our journey into space will go on . "	0.2%	A positive PSA test has to be followed up with a biopsy or other procedures before cancer can be confirmed .	The report states that, before cancer can be confirmed, a positive PSA test must be carried out with a biopsy or other procedures.

Table 14: Additional one-shot paraphrase generation results. 'Ent. score' is the predicted demonstration reference/candidate entailment probability from the classification loss.

#	Type	Demo reference	Demo candidate	Ent. score	Reference	Generated candidate
31	personal	Both bidders agreed to assume about $ 90 million in debt owed on the planes	We agreed to assume about $ 90 million in debt owed on the planes	95.4%	Robert Stewart , a spokesman for Park Place , the parent company of Caesars Palace , said he was surprised by the court 's decision .	Robert Stewart said he was surprised by the court decision.
32	personal	Both bidders agreed to assume about $ 90 million in debt owed on the planes	We agreed to assume about $ 90 million in debt owed on the planes	95.4%	Democrats now hope to increase the value of awards proposed by Hatch and to create a mechanism to ensure the fund remains solvent .	Democrats now hope to increase the value of awards proposed by Hatch.
33	personal	Both bidders agreed to assume about $ 90 million in debt owed on the planes	We agreed to assume about $ 90 million in debt owed on the planes	95.4%	Indonesia 's army has often been accused of human rights abuses during GAM 's battle for independence , charges it has generally denied while accusing the separatists of committing rights violations .	Indonesia has often been accused of human rights abuses during the battle for independence.
34	personal	Both bidders agreed to assume about $ 90 million in debt owed on the planes	We agreed to assume about $ 90 million in debt owed on the planes	95.4%	Progress Software plans to acquire privately held DataDirect Technologies for about $ 88 million in cash , the companies said Friday .	We said we planned to acquire DataDirect Technologies privately for about $ 88 million in cash.
35	personal	Both bidders agreed to assume about $ 90 million in debt owed on the planes	We agreed to assume about $ 90 million in debt owed on the planes	95.4%	A positive PSA test has to be followed up with a biopsy or other procedures before cancer can be confirmed .	We have to follow a positive PSA test with biopsy or other procedures before cancer can be confirmed.

Table 15: Additional one-shot paraphrase generation results. 'Ent. score' is the predicted demonstration reference/candidate entailment probability from the classification loss.

#	Type	Demo reference	Demo candidate	Ent. score	Reference	Generated candidate
36	informal	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both guys agreed to take on like $ 90 million that they owed for some planes	98.8%	Robert Stewart , a spokesman for Park Place , the parent company of Caesars Palace , said he was surprised by the court 's decision .	Robert Stewart, a spokesman for Park Place, the parent company of Caesars Palace, said they were surprised by the court decision.
37	informal	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both guys agreed to take on like $ 90 million that they owed for some planes	98.8%	Democrats now hope to increase the value of awards proposed by Hatch and to create a mechanism to ensure the fund remains solvent .	Democrats now hope to get the value of awards proposed by Hatch and create a mechanism to keep the fund solvent.
38	informal	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both guys agreed to take on like $ 90 million that they owed for some planes	98.8%	Indonesia 's army has often been accused of human rights abuses during GAM 's battle for independence , charges it has generally denied while accusing the separatists of committing rights violations .	Indonesia's army often got accused of human rights abuses at GAM's battle for independence, which they generally denied while accusing the separatists of committing rights violations.
39	informal	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both guys agreed to take on like $ 90 million that they owed for some planes	98.8%	Progress Software plans to acquire privately held DataDirect Technologies for about $ 88 million in cash , the companies said Friday .	They wanted to buy some privately held DataDirect Technologies for about $ 88 million in cash on Friday.
40	informal	Both bidders agreed to assume about $ 90 million in debt owed on the planes	Both guys agreed to take on like $ 90 million that they owed for some planes	98.8%	A positive PSA test has to be followed up with a biopsy or other procedures before cancer can be confirmed .	Some PSA tests need to be followed up with a biopsy or other procedures before they get cancer confirmed.