A Study in Improving BLEU Reference Coverage with Diverse Automatic Paraphrasing

Citation for published version:
Bawden, R, Zhang, B, Yankovskaya, L, Tättar, A & Post, M 2020, A Study in Improving BLEU Reference Coverage with Diverse Automatic Paraphrasing, in Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics (ACL), pp. 918-932, The 2020 Conference on Empirical Methods in Natural Language Processing, Virtual conference, 16/11/20. <https://www.aclweb.org/anthology/2020.findings-emnlp.82>

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Findings of the Association for Computational Linguistics: EMNLP 2020

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
A Study in Improving BLEU Reference Coverage with Diverse Automatic Paraphrasing

Rachel Bawden¹  Biao Zhang¹  Lisa Yankovskaya²  Andre Tättar²  Matt Post³
¹School of Informatics, University of Edinburgh, Scotland
²University of Tartu, Tartu, Estonia
³Johns Hopkins University, Baltimore, Maryland, USA

Abstract

We investigate a long-perceived shortcoming in the typical use of BLEU: its reliance on a single reference. Using modern neural paraphrasing techniques, we study whether automatically generating additional diverse references can provide better coverage of the space of valid translations and thereby improve its correlation with human judgments. Our experiments on the into-English language directions of the WMT19 metrics task (at both the system and sentence level) show that using paraphrased references does generally improve BLEU, and when it does, the more diverse the better. However, we also show that better results could be achieved if those paraphrases were to specifically target the parts of the space most relevant to the MT outputs being evaluated. Moreover, the gains remain slight even when using human paraphrases elicited to maximize diversity, suggesting inherent limitations to BLEU’s capacity to correctly exploit multiple references. Surprisingly, we also find that adequacy appears to be less important, as shown by the high results of a strong sampling approach, which even beats human paraphrases when used with sentence-level BLEU.¹

1 Introduction

There is rarely a single correct way to translate a sentence; work attempting to encode the entire translation space of a sentence suggests there may be billions of valid translations (Dreyer and Marcu, 2012). Despite this, in machine translation (MT), system outputs are usually evaluated against a single reference. This especially affects MT’s dominant metric, BLEU (Papineni et al., 2002), since it is a surface metric that operates on explicit n-gram overlap (see. (1) showing two adequate MT outputs, one with only minimal overlap with the reference):²

(1) Ref: This did not bother anybody .
MT₁: This didn’t bother anybody .
MT₂: Nobody was bothered by this .

Almost since its creation, BLEU’s status as the dominant metric for MT evaluation has been challenged (e.g., Callison-Burch et al. (2006), Mathur et al. (2020)). Such work typically uses only a single reference, however, which is a deficient form of the metric, since one of BLEU’s raisons d’être was to permit the use of multiple references, in a bid to represent “legitimate differences in word choice and word order.” Unfortunately, multiple references are rarely available due to the high cost and effort of producing them. One way to inex- pensively create them is with automatic paraphrasing. This has been tried before (Zhou et al., 2006; Kauchak and Barzilay, 2006), but only recently have paraphrase systems become good enough to generate fluent, high quality sentential paraphrases (with neural MT-style systems). Moreover, it is currently unclear (i) whether adding automatically paraphrased references can provide the diversity needed to better cover the translation space, and (ii) whether this increased coverage overlaps with observed and valid MT outputs, in turn improving BLEU’s correlation with human judgments.

We explore these questions, testing on all into-English directions of the WMT19 metrics shared task (Ma et al., 2019) at the system and segment level. We compare two approaches: (i) generating diverse references with the hope of covering as much of the valid translation space as possible, and (ii) more directly targeting the relevant areas of the translation space by generating paraphrases that contain n-grams selected from the system out-

¹Our code and outputs are available at https://github.com/rbawden/paraphrasing-bleu.

²See Sec. 4.2 and (Papineni et al., 2002, §1.1) for details.
puts. This allows us to compare the effects of diversity against an upper bound that has good coverage. We anchor our study by comparing automatically produced references against human-produced ones on a subset of our data.

Our experiments show that adding paraphrased references rarely hurts BLEU and can provide moderate gains in its correlation with human judgments. Where it does help, the gains are correlated with diversity (and less so adequacy), but see diminishing returns, and fall short of the non-diverse method designed just to increase coverage. Manual paraphrasing does give the best system-level BLEU results, but even these gains are relatively limited, suggesting that diversity alone has its limits in addressing weaknesses of surface-based evaluation metrics like BLEU.

2 Related Work

Paraphrasing for MT evaluation There is a long history of using paraphrasing to overcome the limitations of BLEU-style metrics. Some early approaches rely on external resources (e.g. WordNet) to provide support for synonym matching (Banerjee and Lavie, 2005; Kauchak and Barzilay, 2006; Denkowski and Lavie, 2014). More automatic methods of identifying paraphrases have also been developed. An early example is ParaEval (Zhou et al., 2006), which provides local paraphrase support using paraphrase sets automatically extracted from MT phrase tables. More recently, Apidianaki et al. (2018) exploit contextual word embeddings to build automatic HyTER networks. However they achieve mixed results, particularly when evaluating high performing (neural) models.

The use of MT systems to produce paraphrases has also been studied previously. Albrecht and Hwa (2008) create pseudo-references by using out-of-the-box MT systems and see improved correlations with human judgments, helped by the systems being of better quality than those evaluated. This method was extended by Yoshimura et al. (2019), who filter the pseudo-references for quality. An alternative strategy is to use MT-style systems as paraphrasers, applied to the references. Madnani et al. (2007) show that additional (paraphrased) references, even noisy ones, reduce the number of human references needed to tune an SMT system, without significantly affecting MT quality. However their aim for coverage over quality means that their paraphrases are unlikely to be good enough for use in a final evaluation metric.

Despite the attention afforded to the task, success has been limited by the fact that until recently, there were no good sentence-level paraphrasers. Federmann et al. (2019) showed that neural paraphrasers can now outperform humans for adequacy and cost. Attempts (e.g. Napoles et al., 2016) using earlier MT paradigms were not able to produce fluent output, and publicly available paraphrase datasets have only been recently released (Wieting and Gimpel, 2018; Hu et al., 2019a). Moreover, most works focus on synonym substitution rather than more radical changes in sentence structure, limiting the coverage achieved.

Structurally diverse outputs Diverse generation is important to ensure a wide coverage of possible translations. Diversity, both lexical and structural, has been a major concern of text generation tasks (Colin and Gardent, 2018; Iyyer et al., 2018). State-of-the-art neural MT-style text generation models used for paraphrasing (Prakash et al., 2016; Mallinson et al., 2017) typically suffer from limited diversity in the beam. Techniques such as sampling from the model distribution or from noisy outputs have been proposed to tackle this (Edunov et al., 2018) but can harm output quality.

An effective strategy to encourage structural diversity is to add syntactic information (which can be varied) to the generated text. The constraints can be specified manually, for example by adding a parse tree (Colin and Gardent, 2018; Iyyer et al., 2018) or by specifying more abstract constraints such as rewriting embeddings (Xu et al., 2018). A similar but more flexible approach was adopted more recently by Shu et al. (2019), who augment target training sentences with cluster pseudo-tokens representing the structural signature of the output sentence. When decoding, the top cluster codes are selected automatically using beam search and for each one a different hypothesis is selected. We adopt Shu et al.’s approach here, due to the automatic nature of constraint selection and the flexibility afforded by constraint definition, allowing us to test different types of diversity by varying the type of sentence clustering method.

3 Generating paraphrased references

We look at two ways to produce paraphrases of English references using English–English NMT architectures. The first (Sec. 3.1) aims for maximal lexical and syntactic diversity, in a bid to
better cover the space of valid translations. In contrast, the second (Sec. 3.2) aims to produce paraphrases that target the most relevant areas of the space (i.e. that are as close to the good system outputs as possible). Of course, not all outputs are good, so we attempt to achieve coverage while maintaining adequacy to the original reference by using information from the MT outputs. While less realistic practically, this approach further the study of the relationship between diversity and valid coverage.

3.1 Creating diverse paraphrases
To encourage diverse paraphrases, we use Shu et al.’s (2019) method for diverse MT, which consists in clustering sentences according to their type and training a model to produce outputs corresponding to each type. Applied to our paraphrasing scenario, the methodology is as follows:

1. Cluster target sentences by some property (e.g., semantic, syntactic representation);
2. Assign a code to each cluster and prefix each target sentence in the training data with its code (a pseudo-token), as follows:
   \[
   \text{(2)} \quad \text{\langle cl\_14 \rangle They knew it was dangerous .} \\
   \text{\langle cl\_101 \rangle They had chickens, too .} \\
   \text{\langle cl\_247 \rangle That’s the problem .}
   \]
3. Train an NMT-style paraphrase model using this augmented data;
4. At test time, apply the paraphraser to each reference in the test set; beam search is run for each of the \( n \) most probable sentence codes to produce \( n \) paraphrases per reference.

As in (Shu et al., 2019), we test two different types of diversity: \textit{semantic} using LASER sentential embeddings (Artetxe and Schwenk, 2019) and \textit{syntactic} using a TreeLSTM encoder (Tai et al., 2015). Both methods encode each sentence as a vector, and the vectors are clustered using \( k \)-means into 256 clusters (full details in App. C).

3.2 Output-guided constrained paraphrases
Diversity is good, but even a highly diverse set of references may not necessarily be in the same space as the MT outputs. We attempt to achieve high coverage of the system outputs by using a weak signal from those outputs. The signal we use is \textit{unrewarded} \( n \)-grams from the best systems, which are \( n \)-grams in system outputs absent from the original reference. We identify them as follows. For each sentence in a test set, we find all \( n \)-grams that are (a) not in the reference but (b) present in at least 75% of the system outputs, (c) limited to the top half of systems in the human system-level evaluation (Barrault et al., 2019). Then, for each such \( n \)-gram, we generate one paraphrase of the reference using constrained decoding (Post and Vilar, 2018), with that \( n \)-gram as a constraint. This gives a variable-sized set of paraphrased references for each sentence. In order to limit overfitting to the best systems, we use a cross-validation framework, in which we randomly split the submitted systems into two groups, the first used to compute the \( n \)-gram constraints and the augmented references, and the second half for evaluation. We repeat this ten times and report the average correlation across the splits.

4 Experiments
Our goal is to assess whether we can generate paraphrases that are representative of the translation space and which, when used with BLEU, improve its utility as a metric. We therefore carry out experiments to (i) evaluate the adequacy and diversity of our paraphrases (Sec. 5.2) and (ii) compare the usefulness of all methods in improving BLEU’s correlation with human judgments of MT quality (Sec. 4.1). BLEU is a corpus-level metric, and our primary evaluation is therefore its system-level correlation. However, it is often also used at the segment level (with smoothing to avoid zero counts). It stands to reason that multiple references would be more important at the segment-level, so we also look into the effects of adding paraphrase references for SENTBLEU too.

4.1 Metric evaluation
For each set of extra references, we produce multi-reference BLEU and SENTBLEU metrics, which we use to score all into-English system outputs from the WMT19 news task.\(^3\) We evaluate the

\(^3\)http://statmt.org/wmt19/results.html.
scores as in the metrics task (Ma et al., 2019),
by calculating the correlation with manual direct
assessments (DA) of MT quality (Graham et al.,
2013). System-level scores are evaluated using
Pearson’s r and statistical significance of improve-
ments (against single-reference BLEU) using the
Williams test (Williams, 1959). Segment-level
correlations are calculated using Kendall’s τ (and
significance against single-reference sentBLEU
with bootstrap resampling) on the DA assessments
transformed into relative rankings.

4.2 Baseline and contrastive systems

Our true baselines are case-sensitive corpus BLEU
and sentBLEU, both calculated using sacreBLEU
(Post, 2018) using the standard BLEU formula.
Though likely familiar to the reader, we review it
here. BLEU is computed by averaging modified
n-gram precisions ($p_n$, $n = 1..4$) and multiplying
this product by a brevity penalty (BP), which pe-
nalizes overly short translations and thereby works
to balance precision with recall:

$$\text{BLEU} = \text{BP} \cdot \exp \left( \frac{1}{N} \sum_{n=1}^{N} w_n \log p_n \right)$$

(1)

$$\text{BP} = \begin{cases}
1 & \text{if } c > r \\
\exp \left( 1 - \frac{r/c}{e} \right) & \text{if } c \leq r
\end{cases}$$

(2)

$$p_n = \frac{\sum_{h \in H} \sum_{n_{\text{gram}} \in h} \#(n_{\text{gram}})}{\sum_{h' \in H} \sum_{n_{\text{gram}}' \in h'} \#(n_{\text{gram}}')}$$

(3)

with $c$ and $r$ the lengths of the hypothesis and refer-
ence sets respectively, $H$ is the set of hypo-
thesis translations, $\#(n_{\text{gram}})$ the number of times
$n_{\text{gram}}$ appears in the hypothesis, and $\#(n_{\text{gram}}')$
is the same but clipped to the maximum number of
times it appears in any one reference.

By definition, BLEU is a corpus-level metric,
since the statistics above are computed across sen-
tences over an entire test set. The sentence-level
variant requires a smoothing strategy to counteract
the effect of 0 n-gram precisions, which are more
probable with shorter texts. We use exponential
smoothing. Both baselines use the single provided
reference only. We also compare against several
contrastive paraphrasing approaches: (i) BEAM, which
adds to the provided reference the the n-
best hypotheses in the beam of a baseline para-
phraser, and (ii) SAMPLED, which samples from
the top 80% of the probability mass at each time
step (Edunov et al., 2018). For the sentence en-
coding methods, we also include (iii) RANDOM,
where randomly selected cluster codes are used at
training and test time.

As a topline, we compare against manually
paraphrased references (HUMAN), which we pro-
duce for a subset of 500 sentences from the de–en
test set. Two native English speakers together pro-
duced five paraphrases per reference (alternately
two or three paraphrases). They were instructed
to craft paraphrases that were maximally different
(lexically and syntactically) from both the refer-
ence and the other paraphrases (to which they had
access), without altering the original meaning.

4.3 Paraphrase model training

We train our paraphrasers using data from Para-
bank 2 (Hu et al., 2019b), containing \approx 20M sen-
tences with up to 5 paraphrases each, of which we
use the first paraphrase only. We preprocess by re-
moving duplicate sentences and those longer than
100 words and then segment into subwords using
SentencePiece (Kudo and Richardson, 2018)
(unigram model (Kudo, 2018) of size 16k). The
data splits are created by randomly shuffling the
data and reserving 3k pairs each for dev and test.

For syntactic sentence encoding methods, we use
the Berkeley Parser (Petrov et al., 2006) (internal
tokenisation and prioritizing accuracy) and prune
trees to a depth of 4 for $\approx 6M$ distinct trees.\footnote{Cf. App. A for the number of trees at different depths.}

Paraphrase models are Transformer base mod-
els (Vaswani et al., 2017) (Cf. App. B for details).
All models are trained using the Marian NMT
formula.\footnote{Cf. App. B for the number of trees at different depths.}

As a topline, we compare against manually
paraphrased references (HUMAN), which we pro-
duce for a subset of 500 sentences from the de–en
test set. Two native English speakers together pro-
duced five paraphrases per reference (alternately
two or three paraphrases). They were instructed
to craft paraphrases that were maximally different
(lexically and syntactically) from both the refer-
ence and the other paraphrases (to which they had
access), without altering the original meaning.

4.3 Paraphrase model training

We train our paraphrasers using data from Para-
bank 2 (Hu et al., 2019b), containing \approx 20M sen-
tences with up to 5 paraphrases each, of which we
use the first paraphrase only. We preprocess by re-
moving duplicate sentences and those longer than
100 words and then segment into subwords using
SentencePiece (Kudo and Richardson, 2018)
(unigram model (Kudo, 2018) of size 16k). The
data splits are created by randomly shuffling the
data and reserving 3k pairs each for dev and test.

For syntactic sentence encoding methods, we use
the Berkeley Parser (Petrov et al., 2006) (internal
tokenisation and prioritizing accuracy) and prune
trees to a depth of 4 for $\approx 6M$ distinct trees.\footnote{Cf. App. A for the number of trees at different depths.}

Paraphrase models are Transformer base mod-
els (Vaswani et al., 2017) (Cf. App. B for details).
All models are trained using the Marian NMT
formula.\footnote{Cf. App. B for the number of trees at different depths.}

As a topline, we compare against manually
paraphrased references (HUMAN), which we pro-
duce for a subset of 500 sentences from the de–en
test set. Two native English speakers together pro-
duced five paraphrases per reference (alternately
two or three paraphrases). They were instructed
to craft paraphrases that were maximally different
(lexically and syntactically) from both the refer-
ence and the other paraphrases (to which they had
access), without altering the original meaning.

5 Paraphrase Adequacy and Diversity

5.1 Adequacy

To ensure our automatically produced paraphrases
are of sufficient quality, we first assess their ade-
quacy (i.e., faithfulness to the original mean-
ing). We determine adequacy by manually eval-

By definition, BLEU is a corpus-level metric,
since the statistics above are computed across sen-
tences over an entire test set. The sentence-level
variant requires a smoothing strategy to counteract
the effect of 0 n-gram precisions, which are more
probable with shorter texts. We use exponential
smoothing. Both baselines use the single provided
reference only. We also compare against several
contrastive paraphrasing approaches: (i) BEAM, which
adds to the provided reference the the n-
best hypotheses in the beam of a baseline para-
phraser, and (ii) SAMPLED, which samples from
the top 80% of the probability mass at each time
step (Edunov et al., 2018). For the sentence en-
coding methods, we also include (iii) RANDOM,
using two diversity scores (DS):

We evaluate the diversity of paraphrased references using two diversity scores (DS):

\[ DS_y = \frac{1}{|Y|(|Y| - 1)} \sum_{y \in Y} \sum_{y' \in Y, y' \neq y} 1 - \Delta_y (y, y') \]

where \( Y \) is the set of paraphrases of a sentence produced by a given system, and \( \Delta_y \) calculates the similarity of paraphrases \( y \) and \( y' \). We use two different functions: \( \Delta_{BOW} \) (for lexical similarity) and \( \Delta_{tree} \) (for syntactic similarity). Both give scores between 1 (identical) and 0 (maximally diverse).

### 5.2 Diversity

We evaluate the diversity of paraphrased references using two diversity scores (DS):

\[ DS_{BOW} = \frac{1}{|Y|(|Y| - 1)} \sum_{y \in Y} \sum_{y' \in Y, y' \neq y} 1 - \Delta_{BOW} (y, y') \]

\[ DS_{tree} = \frac{1}{|Y|(|Y| - 1)} \sum_{y \in Y} \sum_{y' \in Y, y' \neq y} 1 - \Delta_{tree} (y, y') \]

\( DS_{BOW} \) is the lexical overlap between the sets of words in two paraphrases. \( \Delta_{BOW} (y, y') \) corresponds to the number of unique words in common between \( y \) and \( y' \), divided by their mean length.

\( DS_{tree} \) uses \( \Delta_{tree} \), the average tree kernel similarity score between paraphrases. We compute tree kernels using the “subset tree” (SST) comparison tree kernel similarity function presented in (Moschitti, 2006, §2.2), with a decay value of \( \lambda = 0.5 \), and excluding leaves (\( \sigma = 0 \)).

| Method | DS_{BOW} | DS_{tree} | BLEU |
|--------|----------|-----------|------|
| RANDOM | 0.10 | 0.01 | 34.8 |
| BEAM   | 0.22 | 0.30 | 37.0 |
| LASER  | 0.24 | 0.33 | 37.5 |
| TREELSTM | 0.28 | 0.47 | 37.7 |
| SAMPLED | 0.41 | 0.56 | 40.1 |
| SAMPLED* Constraints | 0.40 | 0.55 | 47.0 |
| HUMAN | 0.80 | 0.68 | 48.9 |

| Method | DS_{BOW} | DS_{tree} | BLEU |
|--------|----------|-----------|------|
| RANDOM | 0.10 | 0.01 | 34.8 |
| BEAM   | 0.27 | 0.37 | 39.7 |
| LASER  | 0.31 | 0.45 | 41.3 |
| TREELSTM | 0.32 | 0.53 | 41.0 |
| SAMPLED | 0.51 | 0.65 | 47.3 |

Table 2: Diversity scores (DS) of paraphrased references averaged over all into-English test sets, where \( n \) is the number of paraphrases. The final row indicates diversity among MT outputs. * indicates results just for the 500-sentence de–en subset. The final column is the average BLEU score.

The results (Tab. 2) show that all methods other than RANDOM give more diversity than BEAM. Shu et al.’s cluster code method generates diverse paraphrases. As expected, random cluster codes are not helpful, producing mostly identical para-
phrases differing only in the cluster code. Diversity increases for all methods as paraphrases are added. \textsc{TreeLSTM} produces structurally more diverse paraphrases than \textsc{Laser} and has high lexical diversity too, despite codes being entirely syntactic, suggesting that structural diversity leads to varied lexical choices. The most lexically and structurally diverse method (except for \textsc{Human}), is in fact the strong baseline \textsc{Sampled}, which is likely due to the noise added with the method.

The increased diversity is generally reflected by an increase in the average BLEU score (final column of Tab. 2). These higher BLEU scores indicate that the additional paraphrases are better covering the translation space of the MT outputs, but it remains to be seen whether this concerns the space of valid and/or invalid translations. In contrast, some of the diversity makes less of an impact on the BLEU score; the gap in syntactic diversity between \textsc{Laser} and \textsc{TreeLSTM} (+20 references) is not reflected in a similar gap in BLEU score, indicating that this added diversity is not relevant to the evaluation of these specific MT outputs.

6 Metric Correlation Results

The correlation results for each of the metrics (both system- and segment-level) for different numbers of additional references\(^5\) (aggregated full results) are shown in Tab. 3a and Tab. 3b (for the de–en 500-sample subset). We aggregate the main results to make them easier to interpret by averaging over all into-English test sets (the Ave. column) and we also provide the gains for the language pairs that gave the smallest and greatest gains (Min and Max respectively). Full raw results can be found in App. D.

System-level Adding paraphrased references does not significantly hurt performance, and usually improves it; we see small gains for most languages (Ave. column), although the size of the gain varies, and correlations for two directions (fi–en and gu–en) are degraded but non-significantly (shown by the small negative minimum gains).

Fig. 1 (top) shows that for the diverse approaches, the average gain is positively correlated with the method’s diversity: increased diversity does improve coverage of the valid translation space. This positive correlation holds for all directions for which adding paraphrases helps (i.e., all except fi–en and gu–en). For these exceptions, none of the methods significantly improves over the baseline, and \textsc{Random} gives as good if not marginally better results. The constraints approach achieves the highest average gain, suggesting that it is more efficiently targeting the space of valid translations, even though its paraphrases are significantly less diverse (Tab. 2).

Finally, and in spite of these improvements, we note that all systems fall far short of the best WMT19 metrics, shown in the last row. Automatic paraphrases do not seem to address the weakness of BLEU as an automatic metric.

Segment-level Similar results can be seen at the segment level, with most diverse approaches showing improvements over the baseline (this time sent\textsc{BLEU}) and a minority showing nonsignificant deteriorations (i.e., no change). The diversity of the approaches is again positively correlated with the gains seen (Fig. 1, bottom), with the exception of zh–en, for no easily discernable reason.

The best result of the diverse approaches is again achieved by the \textsc{Sampled} baseline.

The constraint-based approach achieves good scores, comparable to \textsc{Sampled}, despite an anomalously poor score for one language pair (for kk–en, with a degradation of 0.097. This approach also had the highest BLEU scores, however, suggesting that the targeted paraphrasing approach here missed its mark.

De–en 500-sentence subset The general pattern shows the same as the averages over all languages in Tab. 3a, with the more diverse methods (especially \textsc{Sampled}) resulting in the greatest gains. The human results also follow this pattern, resulting in the highest gains of all at the system level. Interestingly, the constrained system yields higher average BLEU scores than \textsc{Human} (Tab. 2) yet a comparable system correlation gain, indicating it targets more of the invalid translation space. For this particular subset, the constraints-based approach helps slightly more at the segment level than the system level, even surpassing the human paraphrases in terms of relative gains, despite it having remarkably less diversity.

7 Discussion

Does diversity help? In situations where adding paraphrases helps (which is the case for a majority
Table 3: Absolute gains in correlation (with respect to the true BLEU and sentenceBLEU baseline correlations). Significant gains (except for averages) are marked in bold ($p \leq 0.05$). Full results per language pair are provided in App. D. WMT-19 best refers to the best metric scores from the official shared task (the best metric can be different for each language pair).

![Figure 1: Lexical diversity versus absolute correlation gain at the system level (top) and segment level (bottom) for a variety of paraphrase systems (+2, +5, +10 and +20 references).](image)

(a) Average and minimum and maximum gains over all into-English test sets

(b) 500-sample subset

---

We did not categorize our adequacy judgments, but SAMPLED’s lower adequacy could be caused by (the relatively harmless) deletion of information (anecdotally supported in Tab. 1).
Why are gains only slight? With respect to the SENTBLEU baseline, we calculate the percentage of comparisons for which the decision is improved (the baseline scored the worse translation higher than the better one and the new paraphrase-augmented metric reversed this)\(^7\) and for which the decision is degraded (opposite reversal). The results (Fig. 3) show that although all the systems improve a fair number of comparisons (up to 9.6%), they degrade almost as many. So, while paraphrasing adds references that represent the space of valid translations, references are indeed being added that match with the space of invalid ones too. Interestingly, the same pattern can be seen for human paraphrases, 6.46% of comparisons being degraded vs. 8.30% improved, suggesting that even when gold standard paraphrases are produced, the way in which the references are used by SENTBLEU still rewards some invalid translations, though the balance is shifted slightly in favor of valid translations. This suggests that at least at the segment level, BLEU is a balancing act between rewarding valid translations and avoiding rewarding invalid ones. Some of these effects may be smoothed out in system-level BLEU but there is still likely to be an effect. It is worth noting that for the two languages directions, fi–en and gu–en, for which diversity was negatively correlated with correlation gain (i.e., diversity could be harming performance), the most conservative approach (RANDOM) leads to some of the best results.

What is the effect on individual \(n\)-grams? We study which new \(n\)-grams are being matched by the additional references for the two language directions with the largest system-level correlation gain (ru–en and de–en). For each sentence, we collect and count the \(n\)-grams that were not in the original reference but where in the five paraphrased references of BEAM (missing \(n\)-grams),\(^8\) accumulated across all test set sentences. We also looked at the most frequent \(n\)-grams not found at all, even with the help of the paraphrases (i.e., the unrewarded \(n\)-grams from Sec. 3.2). The results are in Table 4.

Unsurprisingly, most 1-grams are common grammatical words (e.g., a, of, to, in, the) that may be present (or not) in any sentence; it is hard to draw any conclusions. For 4-grams, however, we see some interesting patterns. Present in both lists are acronym variants such as U . S . for ‘United States’ and p . m . for ‘afternoon’ or the 24-hour clock; their presence on both sides indicates success in sometimes grabbing this variant as well as failure to do so consistently. We also see phrasal variants such as , according to and , " he said. These last points corroborate a point made by Freitag et al. (2020, §7.2) that references may omit these common variants. It also suggests a more focused method for generating paraphrases: identify a high-precision set of common variants, and ensure their presence in the set of references, via constrained decoding or other means (in the spirit of Meteor’s (Denkowski and Lavie, 2011) synonym-based matching). We note however, that our paraphrasing methods do seem to contain complementary information as they also tend to improve Meteor too (see results in App. F).

8 Conclusion

We studied the feasibility of using diverse automatic paraphrasing of English references to improve BLEU. Although increased diversity of paraphrases does lead to increased gains in correlation with human judgments at both the system and segment levels, the gains are small and inconsistent. We can do a slightly better job by using

\(^7\)\(^8\)Using sacreBLEU’s default v13a tokenization.
cues from the system outputs themselves to produce paraphrases providing a helpful form of “targeted” diversity. The comparison with manually produced paraphrases shows that there is room for improvement, both in terms of how much diversity is achieved and how much BLEU can be improved. However, the lack of any improvement in some languages points to how hard it is to target this “right kind” of diversity a priori; this, together with the relatively limited gains overall (especially in comparison with the best WMT19 metrics), suggests an intrinsic limit to BLEU’s capacity to handle multiple references.

Acknowledgements

We would like to thank MT Marathon 2019, during which this work was initiated and which also provide us with initial computing resources. This work was supported by funding from the European Union’s Horizon 2020 research and innovation programme under grant agreements No 825299 (GoURMET), 825303 and the UK Engineering and Physical Sciences Research Council (EPSRC) fellowship grant EP/S001271/1 (MTStretch).

References

Joshua Albrecht and Rebecca Hwa. 2008. The role of pseudo references in MT evaluation. In Proceedings of the Third Workshop on Statistical Machine Translation, pages 187–190, Columbus, Ohio. Association for Computational Linguistics.

Marianna Apidianaki, Guillaume Wisniewski, Anne Cocos, and Chris Callison-Burch. 2018. Automated paraphrase lattice creation for HyTER machine translation evaluation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 480–485, New Orleans, Louisiana. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7:597–610.

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

Loïc Barrault, Onofre Bojar, Marta R. Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 1–61, Florence, Italy. Association for Computational Linguistics.

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

Emilie Colin and Claire Gardent. 2018. Generating syntactic paraphrases. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 937–943, Brussels, Belgium. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

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

| N | newly matched ngrams | missing ngrams |
|---|----------------------|----------------|
| 1 | a (494) of (480), (442) to (370) in (364) The (315) the (273) is (204) for (196) has (196) on (193) was (179) have (171) that (166) be (155) at (145) been (140) with (138) and (134) to (921) in (921) on (870) is (802) of (798) a (786) for (568) The (556) with (509) it (508) has (505) are (482) by (480) was (478) have (449) - (443) at (437) as (426) which (386) |

Table 4: Most frequently newly matched and missing n-grams for the de-en and ru-en test sets for BEAM (+5).
Markus Dreyer and Daniel Marcu. 2012. HyTER: Meaning-equivalent semantics for translation evaluation. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 162–171, Montréal, Canada. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Christian Federmann, Oussama Elachqar, and Chris Quirk. 2019. Multilingual whispers: Generating paraphrases with translation. In Proceedings of the 5th Workshop on Noisy User-generated Text (WNUT 2019), pages 17–26, Hong Kong, China. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be Guilty but References are not Innocent. arXiv.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers), pages 200–207, Boston, MA. Association for Machine Translation in the Americas.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019a. Improved lexically constrained decoding for translation and monolingual rewriting. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019b. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 44–54, Hong Kong, China. Association for Computational Linguistics.

Mohit Iyyer, John Wieging, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks.

In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018. Marian: Cost-effective high-quality neural machine translation in C++. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 129–135, Melbourne, Australia. Association for Computational Linguistics.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, pages 455–462, New York City, USA. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentence-Piece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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 the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 62–90, Florence, Italy. Association for Computational Linguistics.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie Dorr. 2007. Using parameter tuning in statistical machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 120–127, Prague, Czech Republic. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation.
metrics. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4984–4997, Online. Association for Computational Linguistics.

Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy. Association for Computational Linguistics.

Courtney Napoles, Chris Callison-Burch, and Matt Post. 2016. Sentential paraphrasing as black-box machine translation. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 62–66, San Diego, California. Association for Computational Linguistics.

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

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 433–440, Sydney, Australia. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Aadiya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashqueel Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual LSTM networks. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2923–2934, Osaka, Japan. The COLING 2016 Organizing Committee.

Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations with sentence codes. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1823–1827, Florence, Italy. Association for Computational Linguistics.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1556–1566, Beijing, China. Association for Computational Linguistics.

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

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

Evan James Williams. 1959. Regression Analysis. Wiley, New York.

Qiongkai Xu, Juyan Zhang, Lizhen Qu, Lexing Xie, and Richard Nock. 2018. D-PAGE: diverse paraphrase generation. CoRR, abs/1808.04364.

Ryoma Yoshimura, Hiroki Shimanaka, Yukio Matsumura, Hayahide Yamagishi, and Mamoru Komachi. 2019. Filtering pseudo-references by paraphrasing for automatic evaluation of machine translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 521–525, Florence, Italy. Association for Computational Linguistics.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating machine translation results with paraphrase support. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 77–84, Sydney, Australia. Association for Computational Linguistics.
A Number of distinct parse trees at different depths

Table 5 shows the number of distinct pruned trees at different depths. We choose a depth of 4 for the syntactic sentence encoding methods in our experiments.

| depth | no leaves | type/token | with leaves |
|-------|-----------|------------|------------|
| 1     | 16        | 0%         | 16         |
| 2     | 207,794   | 1.0%       | 207,794    |
| 3     | 2,158,114 | 11.2%      | 2,629,907  |
| 4     | 6,089,874 | 31.6%      | 10,631,249 |
| 5     | 8,865,720 | 46.1%      | 14,102,645 |
| ∞     | 13,054,272| 68.1%      | 17,362,448 |

Table 5: Number of distinct pruned trees in different depths with and without leaves in the parsed data.

B Paraphraser training details

All paraphrase models are Transformer base models (Vaswani et al., 2017): 6 layers, 8 heads, word embedding dimension of 512, feedforward dimension of 2048. We set dropout to 0.1 and tie all embeddings to the output layer with a shared vocabulary size of 33,152. We use the same vocabulary (including the 256 cluster codes) for all models. We adopt Adam optimisation with a scheduled learning rate (initial $3 \times 10^{-4}$) and mini-batch size of 64. We train each model on 4 GTX Titan X GPUs with a gradient update delay of 2, and select the final model based on validation BLEU.

C Sentence clustering training details

We set $k$ to 256 for $k$-means clustering. We train TREE LSTM sentence encoders using Adagrad with a learning rate of 0.025, weight decay of $10^{-4}$ and batch size of 400 for a maximum of 20 iterations. We set the model size to 256 and limit the maximum number of child nodes to 10.

D Full raw WMT19 results

Table 7 shows the raw correlations of each each paraphrase-augmented BLEU metric on WMT19 (system-level results top and segment-level results bottom). These correspond to the raw scores used to calculate the gains of each method with respect to the true baseline (BLEU or sentenceBLEU) shown in the main results section in Table 3. We indicate the best system from WMT19 as a point of reference.

E Raw results for the de–en 500-sentence subset

| Method          | Correlation System | Correlation Segment |
|-----------------|--------------------|---------------------|
| Baseline (sentence)BLEU | 0.895              | 0.026               |
| Baselines (+5)   | BEAM               | **0.934**           | 0.048               |
|                 | RANDOM             | **0.926**           | 0.043               |
|                 | SAMPLED            | **0.939**           | 0.069               |
| Diversity (+1)   | LASER              | **0.929**           | 0.048               |
|                 | TREE LSTM          | **0.926**           | 0.037               |
| Diversity (+5)   | LASER              | **0.935**           | 0.049               |
|                 | TREE LSTM          | **0.939**           | 0.034               |
| Constraints 4-gram |                  | **0.933**           | 0.064               |
| Human           |                    | **0.948**           | 0.063               |

Table 6: Correlations on the 500-sentence subset.

F Results with the Meteor metric

Although we focus on ways of improving BLEU using paraphrases in this article, as BLEU is the dominant metric, it is also interesting to look at how adding paraphrases could help similar metrics. We apply the same method to improving the Meteor metric (version 1.5) (Denkowski and Lavie, 2014), a metric which already integrates synonym support.

Summarised results (as gains with respect to the single-reference Meteor metric) are shown in Tab. 8 and raw results are shown in Tab. 9 for both system-level and segment-level correlations. We observe that the true baselines (Meteor and sentenceMeteor) are improved in both cases, possibly more so than BLEU and in different ways, showing that the information added by the paraphrases is complementary to the synonym support offered by Meteor.

G Further examples of automatically paraphrased references

We provide additional examples of paraphrased references. As can be seen from Table 10, TREE LSTM gives us more diverse sentences compared to LASER.
### Table 7: WMT19 correlations of paraphrased BLEU for each method against human assessments (# judgments in brackets).

#### (a) Pearson correlations at the system level.

| Approach | Method    | de-en   | fi-en   | gu-en   | kk-en   | lt-en   | ru-en   | zh-en   | Ave   |
|----------|-----------|---------|---------|---------|---------|---------|---------|---------|-------|
| Baseline | BLEU      | 0.890   | 0.985   | 0.799   | 0.943   | 0.969   | 0.862   | 0.888   | 0.905 |
| Paraphrase baselines (+5) | BEAM      | 0.928   | 0.984   | 0.793   | 0.961   | 0.986   | 0.921   | 0.900   | 0.925 |
|          | RANDOM    | 0.916   | 0.986   | 0.805   | 0.957   | 0.983   | 0.908   | 0.898   | 0.922 |
|          | SAMPLED   | 0.937   | 0.984   | 0.798   | 0.966   | 0.989   | 0.929   | 0.902   | 0.929 |
| Diversity (+1) | LASER   | 0.919   | 0.987   | 0.799   | 0.957   | 0.981   | 0.909   | 0.904   | 0.922 |
|          | TREE LSTM | 0.921   | 0.985   | 0.800   | 0.958   | 0.982   | 0.910   | 0.901   | 0.922 |
| Diversity (+5) | LASER   | 0.934   | 0.985   | 0.795   | 0.963   | 0.987   | 0.918   | 0.896   | 0.925 |
|          | TREE LSTM | 0.933   | 0.982   | 0.796   | 0.964   | 0.987   | 0.918   | 0.896   | 0.925 |
| Constraints 4-grams | LASER   | 0.922   | 0.983   | 0.809   | 0.963   | 0.989   | 0.924   | 0.921   | 0.930 |
|          | TREE LSTM | 0.920   | 0.983   | 0.807   | 0.962   | 0.988   | 0.922   | 0.920   | 0.930 |

#### WMT-19 best

| (YIS-1_SRL) | METEOR | (YIS-0) | WMDO | (ESIM) | (YIS-1) | (ESIM) |
|-------------|--------|---------|------|--------|---------|--------|
| 0.950**     | 0.995  | 0.993***| 0.998***| 0.989*  | 0.979**  | 0.988***|

#### (b) Kendall’s τ at the segment level.

| Approach | Method    | de-en   | fi-en   | gu-en   | kk-en   | lt-en   | ru-en   | zh-en   | Ave   |
|----------|-----------|---------|---------|---------|---------|---------|---------|---------|-------|
| Baseline | sentenceBLEU | 0.055   | 0.228   | 0.175   | 0.368   | 0.251   | 0.114   | 0.317   | 0.215 |
| Paraphrase baselines (+5) | BEAM      | 0.061   | 0.250   | 0.189   | 0.371   | 0.281   | 0.129   | 0.317   | 0.228 |
|          | RANDOM    | 0.056   | 0.240   | 0.184   | 0.374   | 0.269   | 0.122   | 0.315   | 0.223 |
|          | SAMPLED   | 0.073   | 0.251   | 0.192   | 0.374   | 0.295   | 0.127   | 0.313   | 0.232 |
| Diversity (+1) | LASER   | 0.061   | 0.244   | 0.187   | 0.368   | 0.276   | 0.121   | 0.314   | 0.225 |
|          | TREE LSTM | 0.061   | 0.242   | 0.185   | 0.383   | 0.278   | 0.123   | 0.315   | 0.227 |
| Diversity (+5) | LASER   | 0.062   | 0.245   | 0.187   | 0.372   | 0.284   | 0.123   | 0.315   | 0.227 |
|          | TREE LSTM | 0.065   | 0.247   | 0.195   | 0.376   | 0.281   | 0.119   | 0.314   | 0.228 |
| Constraints 4-grams | LASER   | 0.090   | 0.242   | 0.161   | 0.271   | 0.323   | 0.122   | 0.314   | 0.218 |
|          | TREE LSTM | 0.093   | 0.242   | 0.161   | 0.271   | 0.323   | 0.122   | 0.314   | 0.218 |

#### WMT-19 best

| (YIS-1_SRL) | (YIS-1) | (YIS-3_SRL) | (YIS-1_SRL) | (YIS-1_SRL) | (YIS-3_SRL) | (YIS-3_SRL) |
|-------------|---------|-------------|-------------|-------------|-------------|-------------|
| 0.199***    | 0.346***| 0.306***    | 0.442***    | 0.380***    | 0.222***    | 0.431***    | 0.333       |

Table 8: Absolute gains in correlation for paraphrased Meteor for WMT19 with respect to the Meteor baseline. Significant gains (except for averages) are marked in bold (p ≤ 0.05).
| Approach                        | Method   | de-en (16) | fi-en (12) | gu-en (12) | kk-en (11) | lt-en (11) | ru-en (14) | zh-en (15) | Ave    |
|--------------------------------|----------|------------|------------|------------|------------|------------|------------|------------|--------|
| Baseline                       | METEOR   | 0.909      | 0.993      | 0.883      | 0.969      | 0.972      | 0.825      | 0.941      | 0.927  |
| Paraphrase baselines (+5)      | BEAM     | 0.927      | 0.994      | 0.887      | 0.976      | **0.983**  | **0.862**  | **0.949**  | 0.940  |
|                                | RANDOM   | 0.920      | 0.994      | 0.889      | 0.974      | **0.981**  | 0.853      | 0.945      | 0.937  |
|                                | SAMPLED  | **0.925**  | 0.995      | 0.891      | **0.978**  | **0.982**  | 0.864      | 0.945      | 0.940  |
| Diversity (+1)                 | LASER    | 0.924      | 0.995      | 0.886      | **0.975**  | **0.979**  | 0.851      | **0.948**  | 0.937  |
|                                | TREE LSTM| **0.923**  | 0.994      | 0.889      | 0.974      | **0.979**  | 0.850      | **0.947**  | 0.937  |
| Diversity (+5)                 | LASER    | 0.932      | 0.995      | 0.890      | **0.978**  | **0.983**  | 0.860      | **0.950**  | 0.941  |
|                                | TREE LSTM| **0.930**  | 0.995      | 0.894      | 0.977      | **0.983**  | 0.864      | **0.950**  | 0.942  |
| Constraints                    | Sentence | 0.922      | 0.990      | 0.910      | 0.983      | 0.988      | 0.775      | 0.949      | 0.931  |
|                                | WMT-19 best | **0.950**  | **0.995**  | **0.993**  | **0.998**  | **0.989**  | **0.979**  | **0.988**  | 0.985  |

(a) Pearson correlations at the system level.

| Approach                        | Method   | de-en (32000) | fi-en (23952) | gu-en (12192) | kk-en (11000) | lt-en (11000) | ru-en (28000) | zh-en (30000) | Ave    |
|--------------------------------|----------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|--------|
| Baseline                       | sentenceMETEOR | 0.061         | 0.243         | 0.197         | 0.356         | 0.275         | 0.145         | 0.351         | 0.233  |
| Paraphrase baselines (+5)      | BEAM     | 0.081         | 0.257         | 0.219         | 0.383         | 0.285         | 0.152         | 0.360         | 0.248  |
|                                | RANDOM   | 0.072         | 0.254         | 0.219         | 0.364         | 0.281         | 0.156         | 0.356         | 0.243  |
|                                | SAMPLED  | 0.080         | 0.262         | 0.228         | 0.375         | 0.292         | 0.160         | 0.360         | 0.251  |
| Diversity (+1)                 | LASER    | 0.079         | 0.258         | 0.209         | 0.370         | 0.283         | 0.150         | 0.359         | 0.244  |
|                                | TREE LSTM| 0.074         | 0.255         | 0.210         | 0.374         | 0.284         | 0.149         | 0.357         | 0.243  |
| Diversity (+5)                 | LASER    | 0.078         | 0.257         | 0.214         | 0.377         | 0.293         | 0.158         | 0.358         | 0.248  |
|                                | TREE LSTM| 0.074         | 0.259         | 0.228         | 0.378         | 0.287         | 0.153         | 0.361         | 0.249  |
| Constraints                    | 4-grams  | 0.098         | 0.237         | 0.193         | 0.272         | 0.318         | 0.145         | 0.351         | 0.230  |
|                                | WMT-19 best | **0.20**      | **0.35**      | **0.31**      | **0.44**      | **0.38**      | **0.22**      | **0.43**      | 0.333  |

(b) Kendall’s τ at the segment level.

Table 9: WMT19 correlations of paraphrased METEOR for each method against human assessments (# judgments in brackets) . Results that are significantly better than the METEOR baseline are indicated as follows (at least $p \leq 0.05$) are marked in bold.
| Reference | Anyway the application phase for each of the 20 properties, which the municipality is offering at better conditions and prices - similar to the earlier local model, has now ended. |
|-----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| LASER     | However, the application phase for each of the 20 properties offered by the municipality at better conditions and prices, similar to the previous local model, is now over. However, the application phase for each of the 20 properties offered by the municipality at better conditions and prices, similar to the previous local model, is now over. |
| TREELSTM  | However, the application phase for each of the 20 properties offered by the municipality at better conditions and prices - similar to the previous local model - is now closed. In any case, the application phase for each of the 20 properties offered by the municipality at better conditions and prices - similar to the previous local model - has now come to an end. |
| Reference | And there is a need for discussion.                                                                                                                                                    |
| LASER     | And we need to discuss it. And we need to have a discussion. And that’s what we need to discuss.                                                                                      |
| TREELSTM  | And there’s a need to discuss it. And we need to discuss it. It is necessary to discuss it.                                                                                           |
| Reference | Why are there increasingly conflicts, despite successful integration?                                                                                                                  |
| LASER     | Why is there ever more conflict, despite successful integration? Why are there more and more conflicts, despite the success of integration?                                                   |
| TREELSTM  | Why is there more and more conflict, despite successful integration? Why are conflicts growing despite successful integration? |
| Reference | Professor Doctor Aladin El-Mafaalani is standing before them.                                                                                                                         |
| LASER     | In front of them is Professor Dr. Aladin El-Mafaalani. Professor Dr. Aladdin El-Mafaalani stands in front of them.                                                                 |
| TREELSTM  | Professor Aladdin El-Mafaalani stands before them. There’s Professor Aladdin El-Mafaalani in front of them.                                                                          |
| Reference | However, the operation took some time.                                                                                                                                                  |
| LASER     | But the surgery took a while. But the operation took a while. But the surgery took a while.                                                                                              |
| TREELSTM  | However, the surgery lasted a while. However, it took some time to operate.                                                                                                             |
| Reference | According to the police, it appears the fire was monitored for a long time, but the embers were inadvertently left behind.                                                              |
| LASER     | According to the police, the fire seems to have been under long surveillance, but the coals were inadvertently left behind. According to the police, the fire seems to have been under long surveillance, but the coals were inadvertently left behind. |
| TREELSTM  | Police say the fire seems to have been under observation for a long time, but the coals are accidentally left behind. The police seem to have been watching the fire for a long time, but the coals were accidentally left behind. |
| Reference | What is the situation on the island now?                                                                                                                                                |
| LASER     | How’s the island now? What’s happening on this island now? What’s the status on the island these days?                                                                             |
| TREELSTM  | What’s it like on the island? What’s going on on the island? So what’s the status on the island?                                                                                         |

Table 10: Top three paraphrases for seven sentences.