Paraphrases as Foreign Languages in Multilingual Neural Machine Translation

Zhong Zhou
Carnegie Mellon University
zhongzhou@cmu.edu

Matthias Sperber
Karlsruhe Institute of Technology
matthias.sperber@kit.edu

Alex Waibel
Carnegie Mellon University
Karlsruhe Institute of Technology
alex@waibel.com

Abstract
Paraphrases, the rewordings of the same semantic meaning, are useful for improving generalization and translation. However, prior works only explore paraphrases at the word or phrase level, not at the sentence or corpus level. Unlike previous works that only explore paraphrases at the word or phrase level, we use different translations of the whole training data that are consistent in structure as paraphrases at the corpus level. We train on parallel paraphrases in multiple languages from various sources. We refer to paraphrases as the different translation versions of the same corpus. We train a unified NMT model on paraphrase-labeled data with a shared attention in the style of multilingual NMT. Similar to multilingual NMT’s objective of translating from any of the $N$ input languages to any of the $M$ output languages (Firat et al., 2016), multi-paraphrase NMT aims to translate as the different translation versions of the same corpus. We train paraphrases in the style of multilingual NMT (Johnson et al., 2017; Ha et al., 2016). Implicit parameter sharing enables multilingual NMT to learn across languages and achieve better generalization (Johnson et al., 2017). Training on closely related languages are shown to improve translation (Zhou et al., 2018). We view paraphrases as an extreme case of closely related languages and view multilingual data as paraphrases in different languages.

Paraphrases can differ randomly or systematically as each carries the translator’s unique style.

1 Introduction
Paraphrases, wordings of texts with preserved semantics, are often used to improve generalization and the sparsity issue in translation (Callison-Burch et al., 2006; Fader et al., 2013; Ganitkevitch et al., 2013; Narayan et al., 2017; Sekizawa et al., 2017). Unlike previous works that use paraphrases at the word/phrase level, we research on different translations of the whole corpus that are consistent in structure as paraphrases at the corpus level;

we refer to paraphrases as the different translation versions of the same corpus. We train paraphrases in the style of multilingual NMT (Johnson et al., 2017; Ha et al., 2016). Implicit parameter sharing enables multilingual NMT to learn across languages and achieve better generalization (Johnson et al., 2017). Training on closely related languages are shown to improve translation (Zhou et al., 2018). We view paraphrases as an extreme case of closely related languages and view multilingual data as paraphrases in different languages. Paraphrases can differ randomly or systematically as each carries the translator’s unique style.

We treat paraphrases as foreign languages, and train a unified NMT model on paraphrase-labeled data with a shared attention in the style of multilingual NMT. Similar to multilingual NMT’s objective of translating from any of the $N$ input languages to any of the $M$ output languages (Firat et al., 2016), multi-paraphrase NMT aims to trans-
late from any of the \( N \) input paraphrases to any of the \( M \) output paraphrases in Figure 1. In Figure 1, we see different expressions of a host showing courtesy to a guest to ask whether sake (a type of alcohol drink that is normally served warm in Asia) needs to be warmed. In Table 6, we show a few examples of parallel paraphrasing data in the Bible corpus. Different translators’ styles give rise to rich parallel paraphrasing data, covering wide range of domains. In Table 7, we also show some paraphrasing examples from the modern poetry dataset, which we are considering for future research.

Indeed, we go beyond the traditional NMT learning of one-to-one mapping between the source and the target text; instead, we exploit the many-to-many mappings between the source and target text through training on paraphrases that are consistent to each other at the corpus level. Our method achieves high translation performance and gives interesting findings. The differences between our work and the prior works are mainly the following.

Unlike previous works that use paraphrases at the word or phrase level, we use paraphrases at the entire corpus level to improve translation performance. We use different translations of the whole training data consistent in structure as paraphrases of the full training data. Unlike most of the multilingual NMT works that uses data from multiple languages, we use paraphrases as foreign languages in a single-source single-target NMT system training only on data from the source and the target languages.

Our main findings in harnessing paraphrases in NMT are the following.

1. Our multi-paraphrase NMT results show significant improvements in BLEU scores over all baselines.
2. Our paraphrase-exploiting NMT uses only two languages, the source and the target languages, and achieves higher BLEUs than the multi-source and multi-target NMT that incorporates more languages.
3. We find that adding the source paraphrases helps better than adding the target paraphrases.
4. We find that adding paraphrases at both the source and the target sides is better than adding at either side.
5. We also find that adding paraphrases with additional multilingual data yields mixed performance; its performance is better than training on language families alone, but is worse than training on both the source and target paraphrases without language families.
6. Adding paraphrases improves the sparsity issue of rare word translation and diversity in lexical choice.

In this paper, we begin with introduction and related work in Section 1 and 2. We introduce our models in Section 3. Finally, we present our results in Section 4 and conclude in Section 5.

2 Related Work

2.1 Paraphrasing

Many works generate and harness paraphrases (Barzilay and McKeown, 2001; Pang et al., 2003; Callison-Burch et al., 2005; Mallinson et al., 2017; Ganitkevitch et al., 2013; Brad and Rebedea, 2017; Quirk et al., 2004; Madnani et al., 2012; Suzuki et al., 2017; Hasan et al., 2016). Some are on question and answer (Fader et al., 2013; Dong et al., 2017), evaluation of translation (Zhou et al., 2006) and more recently NMT (Narayan et al., 2017; Sekizawa et al., 2017). Past research includes paraphrasing unknown words/phrases/sub-sentences (Callison-Burch et al., 2006; Narayan et al., 2017; Sekizawa et al., 2017; Fadaee et al., 2017). These approaches are similar in transforming the difficult sparsity problem of rare words prediction and long sentence translation into a simpler problem with known words and short sentence translation. It is worthwhile to contrast paraphrasing that diversifies data, with knowledge distillation that benefits from making data more consistent (Gu et al., 2017).
Table 1: Comparison of adding source paraphrases and adding target paraphrases. All acronyms including data are explained in Section 4.3.

| Data   | 1   | 6   | 11  | 13  |
|--------|-----|-----|-----|-----|
| Vsrc  | 22.5| 41.4| 48.9| 48.8|
| Vtgt  | 22.5| 40.5| 47.0| 47.4|

Table 2: Comparison of adding a mix of the source paraphrases and the target paraphrases against the baselines. All acronyms including data are explained in Section 4.3.

| data | 1   | 6   | 11  | 16  | 22  | 24  |
|------|-----|-----|-----|-----|-----|-----|
| MrsP | 22.5| 30.8| 29.8| 30.8| 29.3| -   |
| Family | 22.5| 39.3| 45.4| 49.2| 46.6| -   |
| Vmix | 22.5| 44.8| 50.8| 53.3| 55.4| 57.2|
| Vmtf | -   | -   | 49.3| -   | -   | -   |

Our work is different in that we exploit paraphrases at the corpus level, rather than at the word or phrase level.

### 2.2 Multilingual Attentional NMT

Machine polyglotism which trains machines to translate any of the $N$ input languages to any of the $M$ output languages is a new paradigm in multilingual NMT (Firat et al., 2016; Zoph and Knight, 2016; Dong et al., 2015; Gillick et al., 2016; Al-Rfou et al., 2013; Tsvetkov et al., 2016). Many multilingual NMT systems involve multiple encoders and decoders (Ha et al., 2016), and it is hard to combine attention for quadratic language pairs bypassing quadratic attention mechanisms (Firat et al., 2016). An interesting work is training a universal model with a shared attention mechanism with the source and target language labels and Byte-Pair Encoding (BPE) (Johnson et al., 2017; Ha et al., 2016). This method is elegant in its simplicity and its advancement in low-resource language translation and zero-shot translation using pivot-based translation mechanism (Johnson et al., 2017; Firat et al., 2016).

Unlike previous works, our parallelism is across paraphrases, not across languages. In other words, we achieve higher translation performance in the single-source single-target paraphrase-exploiting NMT than that of the multilingual NMT.

### 3 Models

We have four baseline models. Two are single-source single-target attentional NMT models, the other two are multilingual NMT models with a shared attention (Johnson et al., 2017; Ha et al., 2016). In Figure 1, we show an example of multilingual attentional NMT. Translating from all 4 languages to each other, we have 12 translation paths. For each translation path, we label the source sentence with the source and target language tags. Translating from “你喝酒凉了吗?” to “Has your sake turned cold?”, we label the source sentence with _opt_src_zh _opt_tgt_en. More details are in Section 4.

In multi-paraphrase model, all source sentences are labeled with the paraphrase tags. For example, in French-to-English translation, a sentence is tagged with _opt_src_f1 _opt_tgt_e0, denoting that it is translating from version “f1” of French data to version “e0” of English data. In Figure 1, we show 2 Japanese and 2 English paraphrases. Translating from all 4 paraphrases to each other ($N = M = 4$), we have 12 translation paths as $N \times (N - 1) = 12$. For each translation path, we label the source sentence with the source and target paraphrase tags. For example, “お酒冷めましたよね?” to “Has your sake turned cold?”, we label the source sentence with _opt_src_j1 _opt_tgt_e0 in Figure 1. Paraphrases of the same translation path carry the same labels. Our paraphrasing data is at the corpus level, and we train a unified NMT model with a shared attention. Unlike the paraphrasing sentences in Figure 1, we show this example with only one sentence, it is similar when the training data contains many sentences. All sentences in the same paraphrase path share the same labels.

### 4 Experiments and Results

#### 4.1 Data

Our main data is the French-to-English Bible corpus (Mayer and Cysouw, 2014), containing 12 versions of the English Bible and 12 versions of the French Bible. We translate from French to English. Since these 24 translation versions are consistent in structure, we refer to them as paraphrases at corpus level. In our paper, each paraphrase refers to each translation version of whole Bible corpus. To understand our setup, if we use all 12 French paraphrases and all 12 English paraphrases so there are 24 paraphrases in total, i.e., $N = M = 24$, we have 552 translation paths because $N \times (N - 1) = 552$. The original corpus contains missing or extra verses for different

\[1\] We considered the open subtitles with different scripts of the same movie in the same language; they cover many topics, but they are noisy and only differ in interjections. We also considered the poetry dataset where a poem like “If” by Rudyard Kipling is translated many times, by various people into the same language, but the data is small.
| Source Sentence | Machine Translation | Correct Target Translation |
|-----------------|---------------------|-----------------------------|
| Comme de l’eau fraîche pour une personne fatiguée. Ainsi est une bonne nouvelle venant d’une terre lointaine. | As cold waters to a thirsty soul, so is good news from a distant land. | Like cold waters to a weary soul, so is a good report from a far country. |
| Lorsque tu seras invité par quelqu’un à des noces, ne te mets pas à la première place, de peur qu’il n’y ait parmi les invités une personne plus considérable que toi. | When you are invited to one to the wedding, do not be to the first place, lest any one be called greater than you. | When you are invited by anyone to wedding feasts, do not recline at the chief seat lest one more honorable than you be invited by him. |
| Car chaque arbre se connaît à son fruit. On ne cueille pas des figues sur des épines, et l’on ne vendange pas des raisins sur des ronces. | For each tree is known by its own fruit. For from thorns they do not gather figs, nor do they gather grapes from a bramble bush. | For each tree is known from its own fruit. For they do not gather figs from thorns, nor do they gather grapes from a bramble bush. |
| Oui, vous sortirez avec joie. Et vous serez conduits en paix ; Les montagnes et les collines éclateront d’allégresse devant vous. Et tous les arbres de la campagne battront des mains. | When you go out with joy, you shall go in peace; the mountains shall rejoice before you, and the trees of the field shall strike all the trees of the field. | For you shall go out with joy and be led out with peace. The mountains and the hills shall break out into song before you, and all the trees of the field shall clap the palms. |

Table 3: Examples of French-to-English translation trained using 12 French paraphrases and 12 English paraphrases.

paraphrases; we clean and align 24 paraphrases of the Bible corpus and randomly sample the training, validation and test sets according to the 0.75, 0.15, 0.10 ratio. Our training set contains only 23K verses, but is massively parallel across paraphrases.

For all experiments, we choose a specific English corpus as e0 and a specific French corpus as f0 which we evaluate across all experiments to ensure consistency in comparison, and we evaluate all translation performance from f0 to e0.

4.2 Training Parameters

In all our experiments, we use a minibatch size of 64, dropout rate of 0.3, 4 RNN layers of size 1000, a word vector size of 600, a learning rate of 0.8 that decays at the rate of 0.7 if the validation score is not improving or it is past epoch 9 across all LSTM-based experiments. We train 13 epochs. Byte-Pair Encoding (BPE) is used at preprocessing stage (Ha et al., 2016). Our code is built on OpenNMT (Klein et al., 2017) and we evaluate our models using BLEU scores (Papineni et al., 2002), entropy (Shannon, 1951), F-measure and qualitative evaluation.

4.3 Baselines

We introduce a few acronyms for our four baselines to describe the experiments in Table 1, Table 2 and Figure 3. Firstly, we have two single-source single-target attentional NMT models, Single and WMT. Single trains on f0 and e0 and gives a BLEU of 22.5, the starting point for all curves in Figure 3. WMT adds the out-domain WMT’14 French-to-English data on top of f0 and e0; it serves as a weak baseline that helps us to evaluate all experiments’ performance discounting the effect of increasing data.

Moreover, we have two multilingual baselines built on multilingual attentional NMT, Family and Span (Zhou et al., 2018). Family refers to the multilingual baseline by adding one language family at a time, where on top of the French corpus f0 and the English corpus e0, we add up to 20 other European languages. Span refers to the multilingual baseline by adding one span at a time, where a span is a set of languages that contains at least one language from all the families in the data; in other words, span is a sparse representation of all the families. Both Family and Span trains on the Bible in 22 Europeans languages trained using multilingual NMT. Since Span is always suboptimal to Family in our results, we only show numerical results for Family in Table 1 and 2, and we plot both Family and Span in Figure 3. The two multilingual baselines are strong baselines while the fwmt baseline is a weak baseline that helps us to evaluate all experiments’ performance discounting the effect of increasing data. All baseline results are taken from a research work which uses the grid of (1, 6, 11, 16, 22) for the number of languages or equivalent number of unique sentences and we follow the same in Figure 3 (Zhou et al., 2018). All experiments for each grid point carry the same number

---

For multilingual baselines, we use the additional Bible corpus in 22 European languages that are cleaned and aligned to each other.
of unique sentences.

Furthermore, \(V_{src}\) refers to adding more source (English) paraphrases, and \(V_{tgt}\) refers to adding more target (French) paraphrases. \(V_{mix}\) refers to adding both the source and the target paraphrases. \(V_{mf}\) refers to combining \(V_{mix}\) with additional multilingual data; note that only \(V_{mf}\), \(Family\) and \(Span\) use languages other than French and English, all other experiments use only English and French. For the x-axis, data refers to the number of paraphrase corpora for \(V_{src}\), \(V_{tgt}\), \(V_{mix}\); data refers to the number of languages for \(Family\); data refers to and the equivalent number of unique training sentences compared to other training curves for \(WMT\) and \(V_{mf}\).

### 4.4 Results

**Training on paraphrases gives better performance than all baselines:** The translation performance of training on 22 paraphrases, i.e., 11 English paraphrases and 11 French paraphrases, achieves a BLEU score of 55.4, which is +32.9 above the \(Single\) baseline, +8.8 above the \(Family\) baseline, and +26.1 above the \(WMT\) baseline. Note that the \(Family\) baseline uses the grid of (1, 6, 11, 16, 22) for number of languages, we continue to use this grid for our results on number of paraphrases, which explains why we pick 22 as an example here. The highest BLEU 57.2 is achieved when we train on 24 paraphrases, i.e., 12 English paraphrases and 12 French paraphrases.

**Adding the source paraphrases boosts translation performance more than adding the target paraphrases:** The translation performance of adding the source paraphrases is higher than that of adding the target paraphrases. Adding the source paraphrases diversifies the data, exposes the model to more rare words, and enables better generalization. Take the experiments training on 13 paraphrases for example, training on the source (i.e., 12 French paraphrases and the English paraphrase \(e0\)) gives a BLEU score of 48.8, which has a gain of +1.4 over 47.4, the BLEU score of training on the target (i.e., 12 English paraphrases and the French paraphrase \(e0\)). This suggests that adding the source paraphrases is more effective than adding the target paraphrases.

**Adding paraphrases from both sides is better than adding paraphrases from either side:** The curve of adding paraphrases from both the source and the target sides is higher than both the curve of adding the target paraphrases and the curve of adding the source paraphrases. Training on 11 paraphrases from both sides, i.e., a total of 22 paraphrases achieves a BLEU score of 50.8, which is +3.8 higher than that of training on the target side only and +1.9 higher than that of training on the source side only. The advantage of combining both sides is that we can combine paraphrases from both the source and the target to reach 24 paraphrases in total to achieve a BLEU score of 57.2.

**Adding both paraphrases and language families yields mixed performance:** We conduct one more experiment combining the source and target paraphrases together with additional multilingual data. This is the only experiment on paraphrases where we use multilingual data other than only French and English data. The BLEU score is 49.3, higher than training on families alone, in fact, it is higher than training on eight European families altogether. However, it is lower than training on English and French paraphrases alone. Indeed, adding paraphrases as foreign languages is effec-
tive, however, when there is a lack of data, mixing the paraphrases with multilingual data is helpful.

**Adding paraphrases increases entropy and diversity in lexical choice, and improves the sparsity issue of rare words:** We use bootstrap resampling and construct 95% confidence intervals for entropies (Shannon, 1951) of all models of \( V_{mix} \), i.e., models adding paraphrases at both the source and the target sides. We find that the more paraphrases, the higher the entropy, the more diversity in lexical choice as shown in Table 4. From the word F-measure shown in Table 5, we find that the more paraphrases, the better the model handles the sparsity of rare words issue. Adding paraphrases not only achieves much higher BLEU score than the \( WMT \) baseline, but also handles the sparsity issue much better than the \( WMT \) baseline.

**Adding paraphrases helps rhetoric translation and increases expressiveness:** Qualitative evaluation shows many cases where rhetoric translation is improved by training on diverse sets of paraphrases. In Table 3, Paraphrases help NMT to use a more contemporary synonym of “silver”, “money”, which is more direct and easier to understand. Paraphrases simplifies the rhetorical or subtle expressions, for example, our model uses “rejoice” to replace “break out into song”, a personification device of mountains to describe joy, which captures the essence of the meaning being conveyed. However, we also observe that NMT wrongly translates “clap the palm” to “strike”. We find the quality of rhetorical translation ties closely with the diversity of parallel paraphrases data. Indeed, the use of paraphrases to improve rhetoric translation is a good future research question. Please refer to the Table 3 for more qualitative examples.

5 Conclusion

We train on paraphrases as foreign languages in the style of multilingual NMT. Adding paraphrases improves translation quality, the rare word issue, and diversity in lexical choice. Adding the source paraphrases helps more than adding the target ones, while combining both boosts performance further. Adding multilingual data to paraphrases yields mixed performance. We would like to explore the common structure and terminology consistency across paraphrases. As structure and terminology are shared across paraphrases, we are interested in building an explicit representation of the paraphrases and extending our work for better translation, or translation with more explicit and more explainable hidden states, which is very important in all neural systems.

We are interested in broadening our dataset in our future experiments. We hope to use other parallel paraphrasing corpora like the poetry dataset as shown in Table 7. There are very few poems that are translated multiple times into the same language, we therefore need to train on extremely small dataset. Rhetoric in paraphrasing is important in poetry dataset, which again depends on the training paraphrases. The limited data issue is also relevant to the low-resource setting.

We would like to effectively train on extremely small low-resource paraphrasing data. As discussed above about the potential research poetry dataset, dataset with multiple paraphrases is typically small and yet valuable. If we can train using extremely small amount of data, especially in the low-resource scenario, we would exploit the power of multi-paraphrase NMT further.

Cultural-aware paraphrasing and subtle expressions are vital (Levin et al., 1998; Larson, 1984). Rhetoric in paraphrasing is a very important too. In Figure 1, “is your sake warm enough?” in Asian culture is an implicit way of saying “would you like me to warm the sake for you?” We would like to model the culture-specific subtlety through multi-paraphrase training.

**References**

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In Proceedings of the 17th Conference on Computational Natural Language Learning, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Regina Barzilay and Kathleen R McKeown. 2001. Extracting paraphrases from a parallel corpus. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pages 50–57. Association for Computational Linguistics.

Florin Brad and Traian Rebedea. 2017. Neural paraphrase generation using transfer learning. In Proceedings of the 10th International Conference on Natural Language Generation, pages 257–261.

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 255–262. Association for Computational Linguistics.
Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In Proceedings of North American Chapter of the Association for Computational Linguistics on Human Language Technologies, pages 17–24. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, pages 1723–1732.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In Proceedings of the 22nd Conference on Empirical Methods in Natural Language Processing, pages 875–886.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. arXiv preprint arXiv:1705.00440.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1608–1618.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, pages 866–875.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In Proceedings of the 12th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, pages 758–764.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amandeep Subramanya. 2016. Multilingual language processing from bytes. In Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, pages 1296–1306.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. arXiv preprint arXiv:1711.02281.

Thanh-Le Ha, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. arXiv preprint arXiv:1611.04798.

Sadid A Hasan, Bo Liu, Joey Liu, Ashqeul Qadir, Kathy Lee, Vivek Datla, Aadiya Prakash, and Oladimeji Farri. 2016. Neural clinical paraphrase generation with attention. In Proceedings of the Clinical Natural Language Processing Workshop, pages 42–53.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google’s multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. Proceedings of the 55th annual meeting of the Association for Computational Linguistics, System Demonstrations, pages 67–72.

Mildred L Larson. 1984. Meaning-based translation: A guide to cross-language equivalence. University press of America Lanham.

Lori Levin, Donna Gates, Alon Lavie, and Alex Waibel. 1998. An interlingua based on domain actions for machine translation of task-oriented dialogues. In Proceedings of the 5th International Conference on Spoken Language Processing.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In Proceedings of the 11th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, pages 182–190. 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, pages 881–893.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. Oceania, 135(273):40.

Shashi Narayan, Claire Gardent, Shay Cohen, and Anastasia Shimorena. 2017. Split and rephrase. In Proceedings of the 22nd Conference on Empirical Methods in Natural Language Processing, pages 617–627.

Bo Pang, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In Proceedings of North American Chapter of the Association for Computational Linguistics on Human Language Technologies.

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 on association for computational linguistics, pages 311–318. Association for Computational Linguistics.
Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing.

Yuuki Sekizawa, Tomoyuki Kajiwara, and Mamoru Komachi. 2017. Improving Japanese-to-English neural machine translation by paraphrasing the target language. In Proceedings of the 4th Workshop on Asian Translation, pages 64–69.

Claude E Shannon. 1951. Prediction and entropy of printed English. Bell Labs Technical Journal, 30(1):50–64.

Yui Suzuki, Tomoyuki Kajiwara, and Mamoru Komachi. 2017. Building a non-trivial paraphrase corpus using multiple machine translation systems. In Proceedings of ACL 2017, Student Research Workshop, pages 36–42.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, pages 1357–1366.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating machine translation results with paraphrase support. In Proceedings of the 11th Conference on Empirical Methods in Natural Language Processing, pages 77–84. Association for Computational Linguistics.

Zhong Zhou, Matthias Sperber, and Alex Waibel. 2018. Massively parallel cross-lingual learning in low-resource target language translation. In Proceedings of the 3rd Conference on Machine Translation Workshop of the 23rd Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. In Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, pages 30–34.
Appendix A Supplemental Materials

We show a few examples of parallel paraphrasing data in the Bible corpus. We also show some paraphrasing examples from the modern poetry dataset, which we are considering for future research.
If you can fill the unforgiving minute with sixty seconds' worth of distance run, yours is by Rudyard Kipling.

Table 7: Examples of parallel paraphrasing data with German, Chinese, and Portuguese paraphrases of the English poem "If" by Rudyard Kipling.