Paraphrase Generation as Zero-Shot Multilingual Translation: Disentangling Semantic Similarity from Lexical and Syntactic Diversity

Brian Thompson
Johns Hopkins University
brian.thompson@jhu.edu

Matt Post
Johns Hopkins University
post@cs.jhu.edu

Abstract

Recent work has shown that a multilingual neural machine translation (NMT) model can be used to judge how well a sentence paraphrases another sentence in the same language (Thompson and Post, 2020); however, attempting to *generate* paraphrases from such a model using standard beam search produces trivial copies or near copies. We introduce a simple paraphrase generation algorithm which discourages the production of n-grams that are present in the input. Our approach enables paraphrase generation in many languages from a single multilingual NMT model. Furthermore, the amount of lexical diversity between the input and output can be controlled at generation time. We conduct a human evaluation to compare our method to a paraphraser trained on the large English synthetic paraphrase database ParaBank 2 (Hu et al., 2019c) and find that our method produces paraphrases that better preserve meaning and are more gramatical, for the same level of lexical diversity. Additional smaller human assessments demonstrate our approach also works in two non-English languages.

1 Introduction

Paraphrase generation is the task of producing a fluent output sentence which is semantically similar to the input sentence while being syntactically and/or lexically different from it (Bhagat and Hovy, 2013). Paraphrasing has been of longstanding interest in the NLP community (McKeown, 1983) and has been used for data augmentation in question answering (Dong et al., 2017; Gan and Ng, 2019), machine translation (MT) (Hu et al., 2019a; Khayrallah et al., 2020), task oriented dialog (Niu and Bansal, 2018, 2019), and MT metrics (Banerjee and Lavie, 2005; Zhou et al., 2006; Denkowski and Lavie, 2010; Thompson and Post, 2020).

Thompson and Post (2020) recently released the Prism MT metric, which uses a multilingual neural MT (NMT) model as a paraphraser to *score* paraphrastic pairs; they treat paraphrasing as a zero-shot translation task (e.g., “translation” from English to English) and force-decode and score MT system outputs conditioned on their respective human translations. They denote their paraphraser as *lexically/syntactically unbiased* as it does *not* prefer output that differs lexically or syntactically from the input; this is advantageous for an MT metric as it assigns the highest score to an MT output which matches or nearly matches a human reference, but generating from the Prism model using standard beam search produces trivial copies or near copies.

We introduce a simple method to enable paraphrase generation from a multilingual NMT model.[1] Our method discourages the model from producing n-grams that match n-grams in the input sentence. This serves to lexically bias the output away from the input sentence, resulting in non-trivial paraphrases.

When considered together with Prism model of Thompson and Post (2020), our paraphrase generation approach offers several potential advantages over the common technique of training a paraphrase model on synthetic paraphrases generated by translating one side of bitext into the language of the other side (Wieting et al., 2017; Wieting and Gimpel, 2018; Hu et al., 2019c):

- The fluency/semantic similarity vs lexical diversity trade-off can be controlled at generation time.
- The approach works in many languages, with a single model.
- The approach addresses an inherent shortcoming in creating synthetic paraphrases from bi-

[1] We release our code at https://github.com/thompsonb/prism

Proceedings of the 5th Conference on Machine Translation (WMT), pages 561–570
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

text in which ambiguities in one language can create errorful synthetic paraphrases in the other (see §6).

- Separating the fluency and semantic similarity model from the lexical and/or syntactic diversity model allows them to be developed and evaluated with less interdependencies.

We conduct human evaluations to compare our proposed method to a strong English baseline paraphraser trained on the ParaBank 2 dataset (Hu et al., 2019c), which consists of 50 million synthetic examples generated by translating the Czech side of Czech–English bitext into English and pairing it with the original English. We find that our method outperforms this baseline—both in terms of semantic similarity and grammaticality—when our system is adjusted to match the lexical diversity of the baseline. We also present small scale evaluations that suggest our method is effective in other languages.

2 Related Work

Paraphrase Generation Machine translation techniques can be used to train paraphrase models (Quirk et al., 2004). Another method to generate a paraphrase is to translate a text to a different language and then back again (Mallinson et al., 2017). Multiple pivot languages can be used to lessen the effect of inherent ambiguities (Aziz and Specia, 2013), at the expense of complication. Several works have focused on training on paraphrase data, including synthetic data created by starting with bitext and translating one side into the language of the other side to create synthetic paraphrases (Wieting et al., 2017; Wieting and Gimpel, 2018; Hu et al., 2019c). Ideas such as adversarial training (Iyyer et al., 2018), reinforcement learning (Li et al., 2018), and variational autoencoders (Gupta et al., 2018; Chen et al., 2019b) have also been explored in the context of paraphrase generation.

Diversity in Generation Creating paraphrases which differ from their input in non-trivial ways is a challenging problem. Hu et al. (2019c) used constrained decoding (Hokamp and Liu, 2017) in conjunction with a set of constraints (e.g., avoiding certain words which are present in the input) when creating synthetic paraphrases from bitext. Kajiwara (2019) also used hard constraints, but at decoding time. Our work is similar but uses “soft” constraints (i.e., down-weighting tokens which complete n-grams in the input, but not disallowing them all together). Another approach is to control generation with syntactic examples (Iyyer et al., 2018; Chen et al., 2019a) or codes (Shu et al., 2019).

Multilingual NMT Multilingual NMT (Dong et al., 2015) has been shown to enable zero-shot translation—that is, translation between languages pairs not included in training (e.g., translating from Spanish→Arabic at test time when the model was trained on Spanish→English and English→Arabic, but not Spanish→Arabic) (Johnson et al., 2017; Gu et al., 2018; Pham et al., 2019). Zhou et al. (2019) also explored incorporating paraphrase data into training to improve multilingual NMT performance.

Tiedemann and Scherrer (2019) explored using paraphrase recognition to test the semantic abstraction of a fairly small multilingual NMT system trained on Bibles and also demonstrate the model’s ability to paraphrase in English. However, they did not perform a human evaluation of paraphrase quality, and Thompson and Post (2020) found that simply generating via beam search from a multilingual NMT model trained on a large general domain corpus results in trivial copies most of the time. We build upon Tiedemann and Scherrer (2019) by using a larger, general domain model, introducing a novel generation algorithm to produce output with lexical diversity, and performing human evaluations.

Paraphrastic similarity Similarity between intermediate representations produced by multilingual NMT encoders has been used to measure semantic similarity and/or paraphrastic similarity (Schwenk and Douze, 2017; Wieting et al., 2019; Raganato et al., 2019). Similarly, Prism (Thompson and Post, 2020) use a multilingual NMT model as a lexically/syntactically unbiased paraphraser for scoring MT system outputs conditioned on their associated human reference translations. We build on this by introducing a lexical bias away from the input at generation time, enabling the use of a multilingual NMT model as a generative paraphraser.

3 Method

Let x and y be sentences, let $\mathcal{M}(x)$ represent the meaning of x, and let $S(x, y)$ measure the lexical and/or syntactic similarity between the two sentences. Formally, we can state the problem of para-

Algorithm 1 Before paraphrasing a sentence, `buildPenalties()` is called to construct a mapping of word prefixes to subwords that require penalties. Then, `penalize()` is called to modify the model prediction `targetLogProbs` at every decoder timestep.

```
def buildPenalties(source):
  penalties = defaultdict(list)
  for n in [1, 2, 3, 4]:
    for ngram of size n in subwords2words(source):
      prefix, word = ngram[0:-1], ngram[-1]
      for subword in targetVocab:
        if word.lower().startsWith(subword.lower()):
          penalties[prefix].append(subword)
  return penalties

def penalize(history, penalties, targetLogProbs):
  for n in [1, 2, 3, 4]:
    prefix = subwords2words(history)[-(n-1):]
    for subword in penalties[prefix]:
      targetLogProbs[id(subword)] -= alpha * (n ** beta)
```

phrase generation as finding $\hat{y}$:

$$\hat{y} = \underset{y}{\operatorname{argmax}} \left[p(y \mid \mathcal{M}(x)) - \alpha S(x, y) \right] \quad (1)$$

where α controls the semantic similarity and fluency vs lexical and/or syntactic diversity trade-off.

3.1 Lexically/Syntactically Unbiased Paraphraser

The intralingual probability $p(y \mid \mathcal{M}(x))$ can be viewed as a lexically/syntactically unbiased paraphraser. This model is responsible for producing output which is both semantically similar to the input and fluent, but has no notion of lexical and/or syntactic diversity. We use the multilingual NMT system released with Prism to model $p(y \mid \mathcal{M}(x))$.

3.2 Lexical Bias

We choose n-gram overlap as our measure of lexical and/or syntactic similarity $S(x, y)$, and propose a simple n-gram overlap measure that penalizes the production of n-grams matching n-grams in the input sequence to enable the paraphrase generation. Our proposed algorithm begins by constructing a set of all (word) n-grams, $1 \leq n \leq 4$, from the input.[2] At each decoding step, the algorithm checks whether any of the target vocabulary subwords *begin* the last word of an input n-gram.[3] All such subwords are penalized by subtracting αn^{β} from the output log probabilities of the NMT model before selecting candidates to extend the beam, where n is the n-gram length, α is the user-specified trade-off between semantic similarity and lexical diversity, and β is another user-defined hyperparameter.

We experimented with penalizing 1-, 2-, 3-, and 4-grams equally but found it produced disfluent output, as the algorithm tended to avoid all words in the input. The exponential weight allows us to penalize the decoder for producing larger overlapping n-grams more harshly than small ones. All experiments in this work use $\beta = 4$, as this produced output in English which appeared fluent to the authors. Finally, the NMT model's vocabulary contains case variants (e.g., "his" and "His") and we do not want to add variation by trivially changing the case of words, so we penalize all case variants of the next tokens. Pseudocode for our approach is provided in Algorithm 1. Note that this method is much simpler than the method used to generate training data for ParaBank 2, which including hand-written constraints, scoring, filtering, and clustering.

[2] In this work, we assume words are separated by whitespace. For languages which do not denote word boundaries, our method could likely be applied after tokenizing the input, or by simply treating each SentencePiece token as a word.

[3] We apply the penalty at the start of the generation of the last word of an input n-gram so that the decoder is not encouraged to produce an unnatural completion to an already-begun word.

3.3 Diversity Control

The α parameter in Equation 1 provides the user with a knob to control how strongly the output is "pushed" away from the input, in lexical space, during generation. In contrast to positive and negative hard lexical lexical constraints (Hokamp and Liu, 2017; Post and Vilar, 2018; Hu et al., 2019c), our method requires no user-defined constraints, making it simpler and perhaps more language agnostic.[4]

3.4 Development and Evaluation

Paraphrase evaluation is complicated by the fact that many different aspects of paraphrases can be evaluated including semantic similarity between input and output, fluency, grammatical correctness, lexical diversity between input and output, and syntactic diversity between input and output. The relative importance of these aspects is not intuitively obvious and is likely determined by downstream tasks.

Modeling semantic similarity and lexical/syntactic diversity separately has the potential to somewhat lessen the burden of evaluation in several ways:

1. There are several potential ways to automatically evaluate the model $p(y \mid \mathcal{M}(x))$. One option is to evaluate perplexity on a test set consisting of human paraphrases. (Thompson and Post (2020) found that their multilingual NMT model assigned higher probability to both copies of the input and human paraphrases of the input, compared to a model trained on ParaBank 2.) Another option is to test models of $p(y \mid \mathcal{M}(x))$ on pairs of paraphrases where one paraphrase has been deemed to better preserve the semantic meaning of the input. Such datasets already exist, in about a dozen languages, due to the annotation efforts undertaken at the annual WMT evaluations.[5] In other words, we can simply treat a model of $p(y \mid \mathcal{M}(x))$ as an MT metric in order to judge its quality. In other words, we can simply treat a model of $p(y \mid \mathcal{M}(x))$ as an MT metric in order to judge its quality.
2. By applying the lexical/syntactic bias in generation, development of the generation algorithm can be conducted without the time/cost of re-training a model, and multiple generation schemes can be directly compared using the same $p(y \mid \mathcal{M}(x))$ model, such as the freely available Prism model (Thompson and Post, 2020).
3. Being able to control the amount of lexical and/or syntactic diversity at inference time allows for easier comparison with prior paraphrasing work, as the diversity can be adjusted to match that of a prior method. (We employ this approach in §4.3.1.)

4 Experimental Setup

4.1 Primary Model

We use the multilingual NMT model released with Prism (Thompson and Post, 2020), which uses a Transformer (Vaswani et al., 2017) architecture with approximately 750 million parameters. The model was trained in fairseq (Ott et al., 2019). The authors take several steps to encourage the encoder and decoder to be language agnostic, including specifying the target language as the first token in the target, so that the encoder does not know the target language, and training on several datasets that include a large number of different language pairs. The model was trained on several open source datasets including WikiMatrix (Schwenk et al., 2019), Global Voices,[6] EuroParl (Koehn, 2005) SETimes,[7] and United Nations. After filtering, this resulted in approximately 100 million translation pairs and covering 39 languages. The model uses a shared, multilingual vocabulary of 64k SentencePiece tokens (Kudo and Richardson, 2018).

4.2 Baseline Model

As a baseline, we train an English-only paraphraser in fairseq on the ParaBank 2 dataset (Hu et al., 2019c) with approximately 253M parameters and a SentencePiece vocabulary of 16k tokens. We train a Transformer with an 8-layer encoder, 8-layer decoder, 1024 dimensional embeddings, embedding sizes of 1024, feed-forward size of 4096, and 16 attention heads. Dropout is set to 0.3, label smooth-

[4] One concern with hard constraints is that there are sometimes words or phrases (e.g., proper nouns) that should not be paraphrased, as doing so would change the meaning of the sentence. Thus heuristics are often used to determine which words/phrases should be constrained.

[5] In particular, the relative ranking judgements collected through 2016 (Bojar et al., 2016) are probably the most relevant.

[6] `http://casmacat.eu/corpus/global-voices.html`

[7] `http://nlp.ffzg.hr/resources/corpora/setimes/`

Reference	Among other things, the developments in terms of turnover, employment, warehousing and prices are recorded.
α=0.0005	Among other things, developments in terms of turnover, employment, storage and prices are recorded.
α=0.003	Among other things, it records developments in turnover, employment, storage and prices.
α=0.006	Amongst other things, developments regarding turnover, employment, storage and prices were recorded.

Figure 1: Example English paraphrase for the three α values used in this work.

ing to 0.1, and learning rate to 0.0005, and batch size was 31200 tokens. Other parameters match the fairseq defaults. The model trained for approximately 6 weeks (33 epochs) on 4 Nvidia 2080 GPUs.

4.3 Evaluation

We conduct a manual evaluation in English using Mechanical Turk workers and conduct smaller scale manual evaluations in German and Spanish, with the help of colleagues who are native speakers. We perform human evaluations following (Hu et al., 2019b), described in more detail below.

4.3.1 English Evaluation

In this work, we focus on evaluation of semantic similarity, grammatical correctness, and lexical diversity. For the model trained on ParaBank 2, the trade-off between these dimensions is fixed and built into the model. To make a fair comparison, we adjust our overlap penalty (α) such that the output of our method matches the lexical diversity of the model trained on ParaBank 2. Following Hu et al. (2019c), we use uncased BLEU (Papineni et al., 2002), computed between input and output, to estimate the lexical diversity of the paraphraser.

We evaluate in English using Mechanical Turk workers who were selected from a curated list of previously vetted workers. Annotators were presented with a reference sentence and four paraphrases: three paraphrases from our proposed method (at three different operating points) and one from the model trained on ParaBank 2, presented in random order. For each paraphrase, the annotators were asked to (1) rate the paraphrase as (i) grammatical, (ii) having one or two small grammatical errors, or (iii) ungrammatical, and (2) rate the semantic similarity between the input and the paraphrase using an analog slider bar from 1–100. We randomly select 200 sentences from the English side of the WMT19 German–English test set (Barrault et al., 2019) and obtain ratings from three annotators, for each sentence at each paraphrase system/setting combination. Annotators were paid 0.50 USD per HIT.

For our proposed method, we choose three operating points: $\alpha = 0.0005$, α=0.003, and α=0.006 (Figure 1). The middle point of α=0.003 was chosen so as to produce output with the same lexical diversity as the paraphraser trained on ParaBank 2, as described above. We decode with a beam size of 5, using the fairseq defaults.

4.3.2 German & Spanish Evaluation

We also collect human judgments in German and Spanish. We follow the evaluation procedure described above for the English paraphraser except that annotations were done by colleagues who were native speakers in these languages. For Spanish, we used the target side of the WMT 2013 English–Spanish test set (Bojar et al., 2013). For German, we used the target side of the WMT 2019 English–German test set (Barrault et al., 2019). We obtained 50 judgments per set of 3 paraphrases by one German annotator, and 150 judgments per set of 3 paraphrases by three Spanish annotators, both on a random sample of sentences. Multiple paraphrases from our proposed method at different operating points (i.e., different values of α) were shown to the annotator, in random order.

5 Results

5.1 English Results

Human evaluation results in English are shown in Figure 2. We find that α is negatively correlated with grammaticality and semantic similarity between the input and output and positively correlated with lexical diversity of the output with respect to the input, as expected.

We find that at the operating point $\alpha = 0.003$, which was chosen such that our method has the same lexical diversity as the model trained on ParaBank 2, the paraphrases from our method were judged to be both more semantically similar to the input and grammatical (slightly) more often.

5.2 German & Spanish Results

The human evaluation results in German and Spanish, along with English for reference, are shown

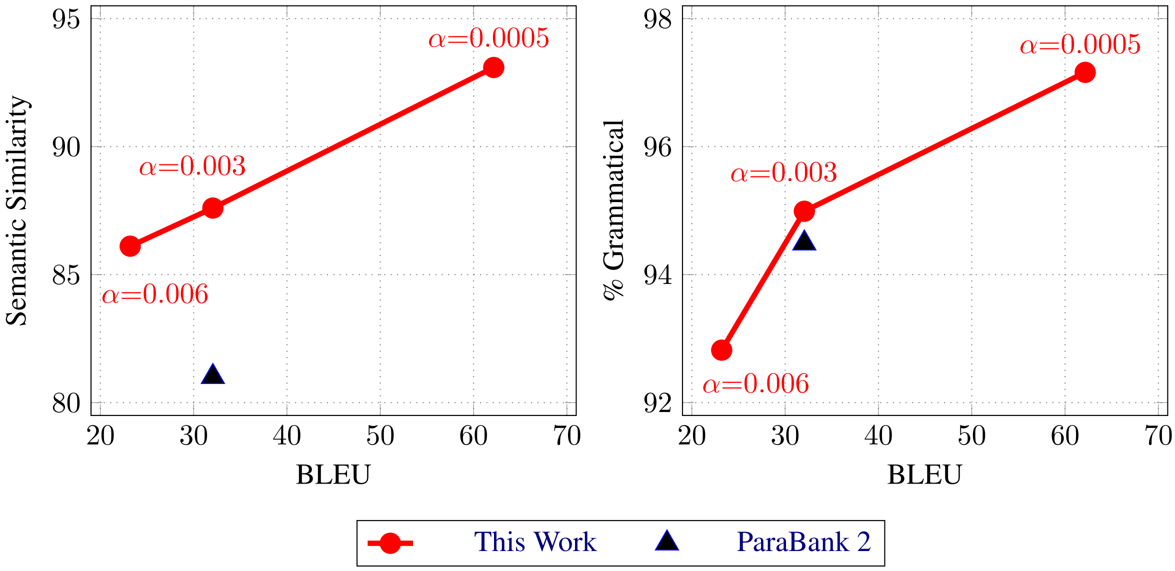

Figure 2: Human judgments of English paraphrases for semantic similarity (rated 1–100) and the percentage of sentences produced which were rated as grammatical, both as a function of lexical/syntactic diversity (measured via uncased BLEU between input and output). We evaluated our generation method at three operating points (α=0.0005, α=0.003, and α=0.006). α=0.003 was chosen to match such that the proposed method had the same diversity as the model trained on Paracrawl2. At that operating point, humans rated output of our method to be more semantically similar to the reference (87.5 vs. 81.0), and grammatical slightly more often (95.0% vs. 94.5%).

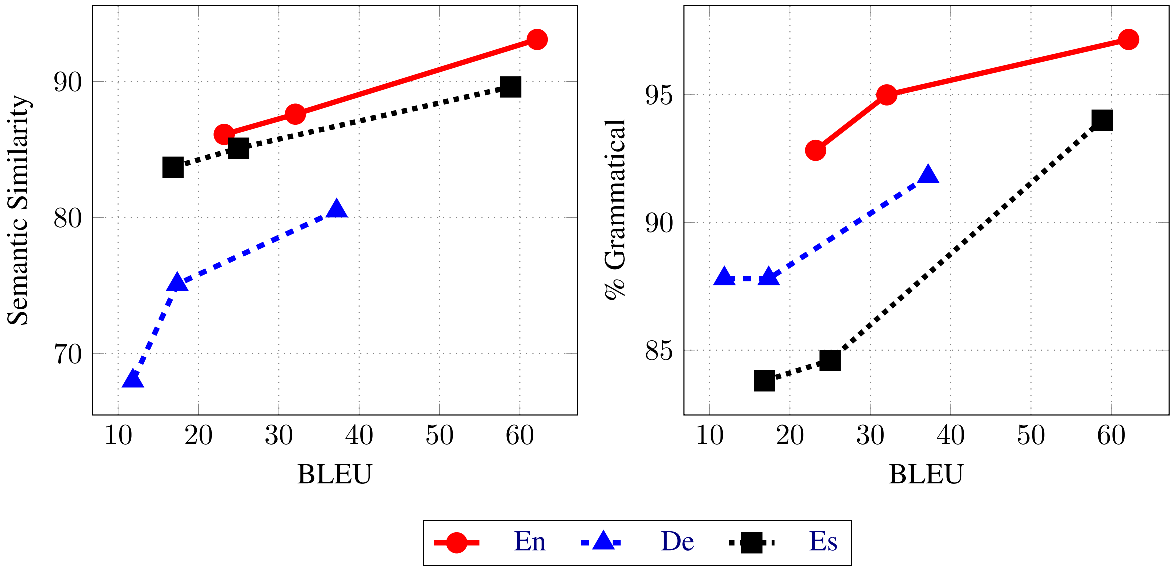

Figure 3: Human judgments of German (De) and Spanish (Es) paraphrases, with English (En) shown for reference, plotted against uncased BLEU computed between the paraphraser input and output. The judgement criteria and α values match English settings. α decreases from left to right in all plots.

in Figure 3. Note that we have no way to normalize between annotators in different languages, thus the results should *not* be used to draw conclusions about the *relative* performance of the paraphraser of these languages. However, we find the trends are similar across all three languages, and that semantic similarity and grammaticality judgements for Spanish and German are both reasonably high.

6 Discussion

We hypothesize that our method outperforms the baseline because it does not suffer from a fundamental shortcoming in creating synthetic paraphrase data from bitext: namely that inherent ambiguities present in one language (but not the other) can cause erroneous synthetic paraphrases in the other language (Aziz and Specia, 2013).

For the sake of discussion we consider gender[8] as an ambiguity. Suppose we create synthetic English paraphrases from Turkish–English data, and our bitext contains the following (valid) sentence pair: ("O mağazaya gitti.", "She went to the store.") Turkish is a gender-neutral language, so when we translate the Turkish side to English it is perfectly valid to translate the sentence to "He went to the store." Pairing the original English translation with the translation results in the synthetic paraphrase example ("She went to the store.", "He went to the store."). Since English is gendered, this results in an invalid synthetic paraphrase.

In contrast, consider what happens if "She went to the store." is paraphrased by our method. First, the sentence is converted to an intermediate representation by the encoder. If the encoder were from an English→Turkish system, it is plausible that the encoder would discard gender information, as it is not needed in the target language. However, our encoder comes from a multilingual system which can produce output in *many* different languages. Thus, as long as the model has seen a sufficient number of training examples between English and at least one other gendered language, we can reasonably expect that the intermediate representation will preserve gender. Thus, when this representation is passed to the decoder and English is requested as the target language, the model should put low probability on any output for which the subject is male.

An alternative way to address pivot language ambiguities is to use multiple pivot languages, as proposed by Aziz and Specia (2013). However, it is not clear how best to extend this idea to neural sequence-to-sequence models, or to a multilingual paraphraser. Combining synthetic paraphrases for training using several different pivot languages would mitigate the errors due to ambiguities from any one pivot language, at the expense of errors due to ambiguities in other pivot languages. To really address such errors would require combining models of different language pairs; see Mallinson et al. (2017) for one such solution.

7 Conclusions

We treat paraphrasing as a zero-shot translation task and present a method to control the lexical diversity of paraphrases generated from a multilingual NMT model, enabling paraphrase generation in many languages. Our approach gives a user fine-grained control over the amount of lexical diversity at generation time, and also allows models and generation algorithms to be developed and evaluated with less interdependencies. There are likely many other ways that the output could be controlled to vary other aspects, such as syntactic diversity (Shu et al., 2019); we would like to explore such methods in future work.

Our work outperforms an English baseline trained on a large synthetic paraphrase dataset (Hu et al., 2019b). This improvement in performance may be because our method does not suffer from the issue that ambiguities in the pivot language used to create synthetic paraphrase data can cause errors in synthetic data. Small experiments indicate our method also performs well in other languages.

Multilingual NMT is an active research area and we are optimistic that this approach will pave the way for even stronger paraphrase generation in the future, as multilingual NMT methods continue to improve and models are publicly released.

Acknowledgments

The authors wish to thank Ben Van Durme for helpful technical discussions, and Carlos Aguirre, Sabrina Mielke, Rachel Wicks, and others for assistance with annotations.

[8] Czech is, of course, gendered, so we would not expect the ParaBank 2 dataset (which was created from Czech–English bitext) to have gender errors. But the logic presented here should generalize to other ambiguities.

References

Wilker Aziz and Lucia Specia. 2013. Multilingual WSD-like constraints for paraphrase extraction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 202–211, Sofia, Bulgaria. Association for Computational Linguistics.

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

Loïc Barrault, Ondřej 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.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019a. Controllable paraphrase generation with a syntactic exemplar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5972–5984, Florence, Italy. Association for Computational Linguistics.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019b. A multi-task approach for disentangling syntax and semantics in sentence representations. 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 2453–2464, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2010. Extending the METEOR machine translation evaluation metric to the phrase level. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 250–253, Los Angeles, California. 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 and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

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

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, Florence, Italy. Association for Computational Linguistics.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. 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 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *AAAI Conference on Artificial Intelligence*.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

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, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019b. ParaBank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of AAAI*.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019c. 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 Wieting, 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.

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

Tomoyuki Kajiwara. 2019. Negative lexically constrained decoding for paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6047–6052, Florence, Italy. Association for Computational Linguistics.

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Taku Kudo and John Richardson. 2018. SentencePiece: 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.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. Paraphrase generation with deep reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3878, Brussels, Belgium. 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.

Kathleen R. McKeown. 1983. Paraphrasing questions using given and new information. *American Journal of Computational Linguistics*, 9(1):1–10.

Tong Niu and Mohit Bansal. 2018. Adversarial over-sensitivity and over-stability strategies for dialogue models. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 486–496, Brussels, Belgium. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2019. Automatically learning data augmentation policies for dialogue tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1317–1323, Hong Kong, China. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. 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.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. 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.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference* *on Empirical Methods in Natural Language Processing*, pages 142–149, Barcelona, Spain. Association for Computational Linguistics.

Alessandro Raganato, Raúl Vázquez, Mathias Creutz, and Jörg Tiedemann. 2019. An evaluation of language-agnostic inner-attention-based representations in machine translation. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 27–32, Florence, Italy. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

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.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2019. Measuring semantic abstraction of multilingual NMT with paraphrase recognition and generation tasks. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 35–42, Minneapolis, USA. Association for Computational Linguistics.

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

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.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Simple and effective paraphrastic similarity from parallel translations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4602–4608, Florence, Italy. Association for Computational Linguistics.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285, Copenhagen, Denmark. Association for Computational Linguistics.

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006. ParaEval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 447–454, New York City, USA. Association for Computational Linguistics.

Zhong Zhou, Matthias Sperber, and Alexander Waibel. 2019. Paraphrases as foreign languages in multilingual neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 113–122, Florence, Italy. Association for Computational Linguistics.