Many thanks for the valuable comments! Our responses follow.

1 Reviewer #4

Q1: About set2seq-common. A: Our main approach is set2seq+BT, which targets majority of the cases where in-domain non-parallel data for paraphrasing is available. set2seq-common is a variant that targets the special cases where in-domain non-parallel data is not available or very limited, such as Twitter, or NMT to low-resource languages. This is why set2seq-common+BT outperforms others for Twitter but underperforms for other datasets in Table 2.
Q2: Compare the proposed framework with other previous models in "Application". A: The sole purpose of our application section is to demonstrate that one possible use of the paraphrases we generate is to augment training data for machine translation for low-resource languages, which is a novel idea. It's by no means our intention to compete with SOTA NMT models.
Q3: Why not fine-tuning? A: For domains with abundant non-parallel data such as Quora or MSCOCO, fine-tuning set2seq-common on the in-domain data will not be better than training set2seq directly on the in-domain data. **We did a follow-up experiments on this for all four datasets in Table 2 and found that ...**
Q4: Using a shared decoder? A: Since set2seq is a lightweight model, and the machine translation (seq2seq) model is complex and heavy, sharing the decoder would make the combined model very expensive to train. We actually made some attempts like this before, but did not get any better results, thus didn't include the idea in the paper.

2 Reviewer #10

Q1: Experiments in NMT is not good enough? A: Please refer to Reviewer #4, Q2.
Q2: About lambda? A: Due to space limitations in the initial submission, we were not able to show all the experiments including those regarding varying lambda. We will definitely put the following results in the final version given an additional page. When λ is close to 0, the result is similar to the reverse translation result. When λ is between 0.4-0.8, the result is stable, and iBLEU is above 14. As λ goes to infinity, the result is slowly approaching that of set2seq.
Q3: Difference with Set Transformer? A: Our set2seq model is essentially a light-weight transformer model without positional encoding. The set transformer, on the other hand, is a transformer without positional encoding, but with additional modifications adapted to computer vision tasks in which input is not a sequence. The use of set2seq model on paraphrase is novel because permutations of words and their variants in the input sequence are analogous to its paraphrases. This dicussion will go into the related work section.

3 Reviewer #29

Q1: Techniques proposed before... As the reviewer rightfully pointed out, our framework is indeed a combination of existing techniques, which works well on paraphrase generation. We thank the reviewer for kind reminder of the references and in the revised version we will definitely cite them and put them in perspective with our framework, given an additional page. One subtle difference between our DAE and the two papers mentioned by reviewer is our unique way of generating the noises which is a set of words.
Q2: Title of paper and back-translation. Thanks for pointing out the difference between "back-translation" and "round-trip translation", which we got confused. We will duely change the title and all references to BT.

4 Reviewer #43

Q1: Setting for lambda? A: Please see Reviewer #10, Q2.
Q2: Number of keywords? A: There are two ways to construct the keyword set (Sec 2.2). First is is to remove all stopwords, which is a non-parametric method. The second takes the top k words with high IDF scores, which is the number of keywords in question. We experimented with both (including varying k) and conclude that removing all stopwords is the best approach overall (see Sec. 3.5 Ablation study)
Q3: Common sense in human evaluation? A: Whether the phrase makes commonsense is implicitly accounted for when scoring accuracy. If the original sentence makes common sense, the correct paraphrase must also conform to common sense.
Q4: Writing needs improvement. A: We will carefully proofread the paper and improve the language through out.

5 Reviewer #97

Q1: Difference between three categories of paraphrasing methods. A: We consider methods that use parallel paraphrase data to be *supervised methods*, methods that do not use parallel paraphrase data but some other form of parallel data (such as translation pairs) to be *distantly supervised methods*, and finally methods that do not use parallel data of any kind to be *unsupervised methods*. Therefore our method belongs to *distantly supervised methods*. Yes, the "back-translation" model was trained on out-of-domain data.
Q2: DNPG outperforms set2seq+BT? A: DNPG is a supervised method, which is listed only for reference. Our method outperforms the DNPG used with domain adaptation.
Q3: About set2seq-common+BT. A: We train the set2seq-common+BT with the English monolingual data from WMT17 (see Section 3.3). This is a large cross-domain dataset, which makes up for the lack of in-domain data for Twitter and benefits the translation task. For other three datasets which have abundant in-domain data, training the set2seq directly on the in-domain data is better than using the general-purpose WMT17 data.
Q4: Problem of computational expensive? A: set2seq is a light-weight while the translation models are relatively heavy. Fortunately the translation model only needs to be trained once. In comparison, Liu et al., ParaNMT and ParaBank also require a translation model, which means their training time will be similar to ours. CGMH and UPSA, however, only need to train a language model using LSTM, so the training time is a few hours.