We thank all reviewers for their thorough reviews and insightful feedback! We are encouraged that they found our work to be a novel [R1], but simple and effective [R4] way to combine two different lines of research on parallel sentence mining and unsupervised machine translation [R1]. We also appreciate that all reviewers found our work well-motivated by an interesting empirical case study [R1, R3, R4], and showed strong results by improving SoTA by significant margins [R1, R3, R4]. We address reviewer comments below and will incorporate all feedback in the final version.

[R4] Novelty compared to [Artetxe et al 2019] First, we thank the reviewer for pointing us to this related work, we will gladly add a reference and discuss it in the final version. However, we would like to clarify how our work is different from it: (1) [Artetxe et al 2019] used cross-lingual word embeddings to build a phrase-based statistical machine translation system, while we use cross-lingual sentence representations to build a neural machine translation system. Therefore, our work is evaluated on tasks such as sentence retrieval, and machine translation, instead of bilingual lexicon induction. (2) Our approach shares the same neural networks architecture for pretraining and downstream tasks, making it easier to finetune for downstream tasks such as mining and translation.

[R4] Novelty compared to other pseudo-parallel sentence mining work CRISS differs from existing pseudo-parallel sentence mining approaches on three important aspects: (1) Compared to supervised approaches such as LASER and [Guo et al 2018], CRISS performs mining with unsupervised sentence representations pretrained from large monolingual data. This enables us to achieve good sentence retrieval performance on very low resource languages such as Kazakh, Nepali, Sinhala, Gujarati. (2) Compared to [Hangya et. al. 2019], we used full sentence representations instead of segment detection through unsupervised word representations. This enabled us to get stronger machine translation results (37.1 BLEU vs 13.07 BLEU on WMT16 de-en). (3) Our case study demonstrated that fine-tuning even on a single language pair significantly improves the quality of retrieval on all language pairs. As mentioned by R1, this is an important new empirical finding that enabled us to iteratively self-improve the model for both mining and translation. We will add an additional related work subsection to discuss the above mentioned methods.

[R4] Comparison to mBART as a strong starting point While we agree that mBART is a strong starting point, all of our results in unsupervised machine translation, sentence retrieval, and supervised machine translation are compared to mBART itself (as well as other pretraining techniques). We also included results after each iteration to show the quality improving after each step, so we believe we showed clear benefits from the iterative mining-training procedure.

[R4] Applying CRISS-style finetuning on other pretraining techniques We agree that CRISS-style finetuning can be applied to other pretraining techniques such as XLM-R/MASS, and we welcome future work in this area. For this paper, we chose to start with mBART since it compared favorably with other methods on machine translation downstream tasks as well as due to page limit.

[R4] Limit in the number of languages We agree that translation for low-resource languages is far from solved, and will clarify in the broader impact section that even though this work contributes to low-resource language translation, more efforts are needed by the community. CRISS' contribution to low resource translation is exemplified by our experiments on 25 languages used in mBART which contains low resource languages such as Nepali and Sinhala in Table 1 and Table 3. We will continue to explore more languages in our future work.

[R4] Evaluation of unsupervised machine translation We fully agree with the reviewer that unsupervised machine translation should be evaluated on low-resource languages. We included results on En-De and En-Fr so that we can make a fair comparison with previous work on unsupervised machine translation, but we also reported results on many low-resource languages, such as the Flores test set (Ne, Si) (Table 1), and WMT 2019 (Gu, Kk) (Table 3 of supplementary materials)

[R4] Starting with bilingual pretrained mBART We agree with the reviewer that the results of training CRISS starting from mBART-2 En-Ro would be instructive for the reader. We will include this experiment in the final version.

[R1, R4] Additional ablation studies on number of languages and scale We had ablation studies comparing bilingual finetuning versus multilingual finetuning (Figure 4,5), and comparing between different numbers of pivot languages (Figure 6,7). In the final version, we will also include an additional ablation study on how the size of monolingual data used in mining affects unsupervised machine translation performance.

[R4] Combination with backtranslation We tried finetuning CRISS further using backtranslation, but weren't able to achieve better performance. We conjecture that the mined data generated from previous iterations made the additional backtranslation data somewhat redundant/less effective.