We thank the reviewers for their thorough and insightful reviews. We first respond to a common question regarding the impact of pretraining in the supervised MT setting, then provide individual answers:

Supervised MT pretraining Reviewers 1 and 2 asked about the impact of pretraining in supervised MT on other language pairs than English-Romanian. We recently ran similar experiments on English-German and English-French. We observed that on En-De (≈6M parallel sentences), pretraining improved the performance by 1 / 1.5 BLEU. On En-Fr (≈40M parallel sentences), however, pretraining did not help. On En-Ro (≈600k parallel sentences) the improvement was of 7 BLEU points. This suggests that pretraining is the most effective in low-resource scenarios. Also, in all settings (En-Fr, En-De, En-Ro) we observed that the convergence when starting from a pretrained model is extremely fast (less than 1 epoch to reach the final performance on En-Fr). However, the overall training time is longer if we include the pretraining time. We will add these results to the paper.

Reviewer 1

- For all language modeling experiments, we used the same architecture as for the NMT experiments: 6 layers, dimension 1024, 8 heads. For each configuration (Ne, Ne+En, Ne+Hi and Ne+Hi+En), we independently tuned the dropout, attention dropout, and optimizer learning rate. Overall, each configuration was given the same amount of hyper-parameters fine-tuning. We will detail this in the paper.
- The missing citation to ELMo was clearly an oversight on our part. We will add it to the paper.
- XNLI only provides a ground-truth training set in English, so we fine-tune on this English training set and evaluate on other languages at test time. In the "translate-train" baseline, however, we fine-tune on each of the translated training sets.
- Transformers are indeed SOTA on language modeling. We will clarify this in the paper.

Reviewer 2

- We use language tags to specify the current language (so if the encoder or decoder receives a German sentence, a German embedding <DE> is added at each time step). Unlike Johnson et al., we do not need to add a <TO_EN> token to the encoder input since the decoder already receives an embedding indicating the target language. Note that adding the language embedding to the encoder is actually optional, but it allows us to initialize the encoder and the decoder with a same pretrained model trained with the language embedding.
- Since we use a language model to pretrain the encoder and the decoder, we cannot pretrain the source-attention parameters that are specific to the decoder. As a result, we simply let these parameters randomly initialized (we use the default He initialization of linear layers).

Reviewer 3

- We tried 3 different strategies to leverage parallel data: align hidden representations of parallel sentences (as in the XNLI paper), predict whether pairs of sentences are mutual translations of each other, and the TLM objective presented in the paper. Overall, the TLM approach gave the best results, but we agree that comparing with other strategies is interesting, and we will develop this in the updated version of the paper.
- The approach of Artetxe et al. would probably benefit from cross-lingual pretraining. Indeed, their method relies on an unsupervised PBSMT training to provide back-parallel sentences, and the NMT model used in a second part could clearly be pretrained and we expect that it would improve the results even further. However, this pipeline requires a significant engineering effort (especially for the PBSMT part) and we did not try it as it is slightly beyond the scope of this paper.
- We agree that in the UNMT setting, results on distant language pairs would also be interesting. However, the majority of existing studies in UNMT focused on En-Fr, En-De and En-Ro, so we also used these language pairs for easier comparison with existing approaches.
- We believe that there are a few non-optimal strategies in the m-BERT approach that may explain the differences. First, we observed that the "next sentence prediction task" was hurting performance, which is in line with other recent studies that focused on the monolingual setting. Also, we create batches of continuous sentence streams, which results in larger attention spans than when training on sentence pairs. Overall, our training is simpler, and we observed improvements both in the monolingual setting (on GLUE tasks), and on cross-lingual tasks.
- Our work is the first to show the importance of pretraining for generation in a multilingual MT setting, and we also demonstrate that cross-lingual pretraining at the encoder/decoder level - instead of just the token embeddings - is critical for UNMT. These contributions are in our opinion quite novel, on top of the TLM approach we introduce for cross-lingual classification. Cross-lingual pretraining is maybe a natural extension, but the fact that it provides substantial improvements was not obvious in our opinion.