Improving the Lexical Ability of Pretrained Language Models for Unsupervised Neural Machine Translation

Alexandra Chronopoulou, Dario Stojanovski, Alexander Fraser

Center for Information and Language Processing, LMU Munich, Germany
{achron, stojanovski, fraser}@cis.lmu.de

Abstract

Successful methods for unsupervised neural machine translation (UNMT) employ cross-lingual pretraining via self-supervision, often in the form of a masked language modeling or a sequence generation task, which requires the model to align the lexical- and high-level representations of the two languages. While cross-lingual pretraining works for similar languages with abundant corpora, it performs poorly in low-resource and distant languages. Previous research has shown that this is because the representations are not sufficiently aligned. In this paper, we enhance the bilingual masked language model pretraining with lexical-level information by using type-level cross-lingual subword embeddings. Empirical results demonstrate improved performance both on UNMT (up to 4.5 BLEU) and bilingual lexicon induction using our method compared to a UNMT baseline.

1 Introduction

UNMT is an effective approach for translation without parallel data. Early approaches transfer information from static pretrained cross-lingual embeddings to the encoder-decoder model to provide an implicit bilingual signal (Lample et al., 2018a; Artetxe et al., 2018c). Lample and Conneau (2019) suggest to instead pretrain a bilingual language model (XLM) and use it to initialize UNMT, as it can successfully encode higher-level text representations. This approach largely improves translation scores for language pairs with plentiful monolingual data. However, while UNMT is effective for high-resource languages, it yields poor results when one of the two languages is low-resource (Guzmán et al., 2019). Marchisio et al. (2020) show that there is a strong correlation between bilingual lexicon induction (BLI) and final translation performance when using pretrained cross-lingual embeddings, converted to phrase-tables, as initialization of a UNMT model (Artetxe et al., 2019). Vulić et al. (2020) observe that static cross-lingual embeddings achieve higher BLI scores compared to multilingual language models (LMs), meaning that they obtain a better lexical-level alignment. Since bilingual LM pretraining is an effective form of initializing a UNMT model, improving the overall representation of the masked language model (MLM) is essential to obtaining a higher translation performance.

In this paper, we propose a new method to enhance the embedding alignment of a bilingual language model, entitled *lexically aligned* MLM, that serves as initialization for UNMT. Specifically, we learn type-level embeddings separately for the two languages of interest. We map these monolingual embeddings to a common space and use them to initialize the embedding layer of an MLM. Then, we train the MLM on both languages. Finally, we transfer the trained model to the encoder and decoder of an NMT system. We train the NMT system in an unsupervised way. We outperform a UNMT baseline and demonstrate the importance of cross-lingual mapping of token-level representations. We also conduct an analysis to investigate the correlation between BLI, 1-gram precision and translation results. We finally investigate whether cross-lingual embeddings should be updated or not during the MLM training process, in order to preserve lexical-level information useful for UNMT. We make the code used for this paper publicly available[1].

2 Proposed Approach

Our approach has three distinct steps, which are described in the following subsections.

2.1 *VecMap* Embeddings

Initially, we split the monolingual data from both languages using BPE tokenization (Sennrich et al.,

[1] https://github.com/alexandra-chron/lexical_xlm_relm

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 173–180
June 6–11, 2021.

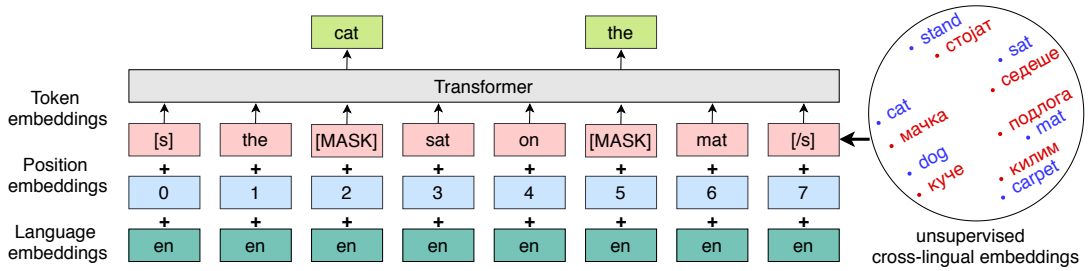

Figure 1: Lexically aligned cross-lingual masked language model.

2016b). We build *subword* monolingual embeddings with *fastText* (Bojanowski et al., 2017). Then, we map the monolingual embeddings of the two languages to a shared space, using *VecMap* (Artetxe et al., 2018a), with identical tokens occurring in both languages serving as the initial seed dictionary, as we do not have any bilingual signal. This is different from the original *VecMap* approach, which operates at the *word level*. We use the mapped embeddings of the two languages to initialize the embedding layer of a Transformer-based encoder (Vaswani et al., 2017).

2.2 Masked Language Model Training

We initialize the token embedding layer of the MLM Transformer encoder with pretrained *VecMap* embeddings, which provide an informative mapping, i.e., cross-lingual lexical representations. We train the model on data from both languages, using masked language modeling. Training a masked language model enhances the cross-lingual signal by encoding contextual representations. This step is illustrated in Figure 1.

2.3 Unsupervised NMT

Finally, we transfer the MLM-trained encoder Transformer to an encoder-decoder translation model. We note that the encoder-decoder attention of the Transformer is randomly initialized. We then train the model for NMT in an unsupervised way, using denoising auto-encoding (Vincent et al., 2008) and back-translation (Sennrich et al., 2016a), which is performed in an online manner. This follows work by Artetxe et al. (2018b); Lample et al. (2018a,c).

3 Experiments

Datasets. We conduct experiments on English-Macedonian (En-Mk) and English-Albanian (En-Sq), as Mk, Sq are low-resource languages, where lexical-level alignment can be most beneficial. We use 3K randomly sampled sentences of SETIMES (Tiedemann, 2012) as validation/test sets. We also use 68M En sentences from NewsCrawl. For Sq and Mk we use all the CommonCrawl corpora from Ortiz Suárez et al. (2019), which are 4M Sq and 2.4M Mk sentences.

Baseline. We use a method that relies on cross-lingual language model pretraining, namely XLM (Lample and Conneau, 2019). This approach trains a bilingual MLM separately for En-Mk and En-Sq, which is used to initialize the encoder-decoder of the corresponding NMT system. Each system is then trained in an unsupervised way.

Comparison to state-of-the-art. We apply our proposed approach to RE-LM (Chronopoulou et al., 2020), a state-of-the-art approach for low-resource UNMT. This method trains a monolingual En MLM model (*monolingual pretraining step*). Upon convergence, a vocabulary extension method is used, that randomly initializes the newly added vocabulary items. Then, the MLM is fine-tuned to the two languages (*MLM fine-tuning step*) and used to initialize an encoder-decoder model. This method outperforms XLM on low-resource scenarios.

Lexically aligned language models. When applied to the baseline, our method initializes the embedding layer of XLM with unsupervised cross-lingual embeddings. Then, we train XLM on the two languages of interest with a masked language modeling objective. Upon convergence, we transfer it to the encoder and decoder of an NMT model, which is trained in an unsupervised way.

In the case of RE-LM, our method is applied to the *MLM fine-tuning step*. Instead of randomly initializing the new embedding vectors added in this step, we use pretrained unsupervised cross-lingual embeddings. We obtain them by applying *VecMap* to *fastText* pretrained Albanian/Macedonian embeddings and the English MLM token-level embeddings. Then, the MLM is fine-tuned on both languages. Finally, it is used to initialize an encoder-decoder NMT model.

	Mk→En		En→Mk		Sq→En		En→Sq	
	BLEU ↑	CHRF1 ↑	BLEU ↑	CHRF1 ↑	BLEU ↑	CHRF1 ↑	BLEU ↑	CHRF1 ↑
XLM	20.7	48.5	19.8	42.4	31.1	56.8	31.3	56.2
lexically aligned XLM	25.2	49.9	22.9	43.1	32.8	58.2	33.5	56.8
RE-LM	25.0	51.1	23.9	45.8	30.1	55.8	32.2	56.4
lexically aligned RE-LM	25.3	51.5	25.6	47.6	30.5	56.0	32.9	56.7

Table 1: UNMT results for translations to and from English. The first column indicates the pretraining method used. The scores presented are significantly different ($p < 0.05$) from the respective baseline. CHRF1 refers to character n-gram F1 score (Popović, 2015). The models in italics are ours.

Unsupervised *VecMap* bilingual embeddings. We build monolingual embeddings with the *fastText* skip-gram model with 1024 dimensions, using our BPE-split (Sennrich et al., 2016b) monolingual corpora. We map them to a shared space, using *VecMap* with identical tokens. We concatenate the aligned embeddings of the two languages and use them to initialize the embedding layer of XLM, or the new vocabulary items of RE-LM.

Preprocessing. We tokenize the monolingual data and validation/test sets using Moses (Koehn et al., 2006). For XLM (Lample and Conneau, 2019), we use BPE splitting with 32K operations jointly learned on both languages. For RE-LM (Chronopoulou et al., 2020), we learn 32K BPEs on En for pretraining, and then 32K BPEs on both languages for the fine-tuning and UNMT steps. The BPE merges are learned on a subset of the En corpus and the full Sq or Mk corpus.

Model hyperparameters. We use a Transformer architecture for both the baselines and UNMT models, using the same hyperparameters as XLM. For the encoder Transformer used for masked language modeling, the embedding and model size is 1024 and the number of attention heads is 8. The encoder Transformer has 6 layers, while the NMT model is a 6-layer encoder/decoder Transformer. The learning rate is set to 10^{-4} for XLM and UNMT. We train the models on 8 NVIDIA GTX 11 GB GPUs. To be comparable with RE-LM, we retrain it on 8 GPUs, as that work reports UNMT results with only 1 GPU. The per-GPU batch size is 32 during XLM and 26 during UNMT. Our models are built on the publicly available XLM and RE-LM codebases. We generate final translations with beam search of size 5 and we evaluate with SacreBLEU[2] (Post, 2018).

4 Results

Table 1 shows the results of our approach compared to two pretraining approaches that rely on MLM training, namely XLM and RE-LM. The lexically aligned XLM improves translation results over the baseline XLM model. We obtain substantial improvements on En-Sq in both directions, of at most 2.2 BLEU and 1.4 CHRF1, while on En-Mk, we get an even larger performance boost of up to 4.5 points in terms of BLEU and 1.4 in terms of CHRF1. Our lexically aligned RE-LM also consistently outperforms RE-LM, most notably in the En→Mk direction, by up to 1.7 BLEU. At the same time, CHRF1 score improves by up to 1.8 points using the lexically aligned pretraining approach compared to RE-LM.

In the case of XLM, the effect of cross-lingual lexical alignment is more evident for En-Mk, as Mk is less similar to En, compared to Sq. This is mainly the case because the two languages use a different alphabet (Latin for En and Cyrillic for Mk). This is also true for RE-LM when translating out of En, showing that enhancing the *fine-tuning step* of MLM with pretrained embeddings is helpful and improves the final UNMT performance.

In general, our method provides better alignment of the lexical-level representations of the MLM, thanks to the transferred *VecMap* embeddings. We hypothesize that static cross-lingual embeddings enhance the knowledge that a cross-lingual masked language model obtains during training. As a result, using them to bootstrap the pretraining procedure improves the ability of the model to map the distributions of the two languages and yields higher translation scores. Overall, our approach consistently outperforms two pretraining models for UNMT, providing for the highest BLEU and CHRF1 scores on all translation directions.

5 Analysis

We conduct an analysis to assess the contribution of lexical-level alignment in the MLM training. We present Bilingual Lexicon Induction (BLI) and

[2]Signature "BLEU+c.mixed+#.1+s.exp+tok.13a+v.1.4.9"

	En-Mk		En-Sq	
	NN	CSLS	NN	CSLS
XLM	6.3	6.5	43.0	40.7
lexically aligned XLM	**15.5**	**16.5**	**51.6**	**50.6**
RE-LM	29.8	16.1	52.0	35.9
lexically aligned RE-LM	**32.0**	**17.2**	**53.0**	**36.9**

Table 2: P@5 results for the BLI task on the MUSE (Lample et al., 2018b) dictionaries. We evaluate the alignment of the embedding layer of each trained MLM.

	En-Mk		En-Sq	
	←	→	←	→
XLM	53.1	41.4	62.1	60.4
lexically aligned XLM	**56.0**	**51.8**	**63.6**	**61.5**
RE-LM	56.0	52.8	61.6	61.2
lexically aligned RE-LM	**56.6**	**53.9**	**62.0**	**61.7**

Table 3: BLEU 1-gram precision scores.

BLEU 1-gram precision scores. We also investigate the best method to leverage pretrained cross-lingual embeddings during MLM training, in terms of final UNMT performance.

Bilingual Lexicon Induction (BLI). We use BLI, a standard way of evaluating lexical quality of embedding representations (Gouws et al., 2015; Ruder et al., 2019), to explore the effect of the alignment of our method. We compare the BLI score of different cross-lingual pretrained language models. We report precision@5 (P@5) using nearest neighbors (NN) and cross-lingual semantic similarity (CSLS). The results are presented in Table 2. We use the embedding layer of each MLM for this task. We also experimented with averages over different layers, but noticed the same trend in terms of BLI scores. We obtain word-level representations by averaging over the corresponding subword embeddings. It is worth noting that we compute the type-level representation of each vocabulary word in isolation, similar to Vulić et al. (2020).

In Table 2, we observe that lexical alignment is more beneficial for En-Mk. This can be explained by the limited vocabulary overlap of the two languages, which does not provide sufficient cross-lingual signal for the training of MLM. By contrast, initializing an MLM with pretrained embeddings largely improves performance, even for a higher-performing model, such as RE-LM. In En-Sq, the effect of our approach is smaller yet consistent. This can be attributed to the fact that the two languages use the same script.

Overall, our method enhances the lexical-level information captured by pretrained MLMs, as shown empirically. This is consistent with our intuition that cross-lingual embeddings capture a bilingual signal that can benefit MLM representations.

1-gram precision scores. To examine whether the improved translation performance is a result of the lexical-level information provided by static embeddings, we present 1-gram precision scores in Table 3, as they can be directly attributed to lexical alignment. The biggest performance gains (up to +10.4) are obtained when the proposed approach is applied to XLM. This correlates with the BLEU scores of Table 1. Moreover, the En-Mk language pair benefits more than En-Sq from the lexical-level alignment both in terms of 1-gram precision and BLEU. These results show that the improved BLEU scores can be attributed to the enhanced lexical representations.

Alignment Method	En-Mk		En-Sq	
	←	→	←	→
lexically aligned MLM				
frozen embeddings	24.7	22.1	31.0	32.1
fine-tuned embeddings (ours)	**25.2**	**22.9**	**32.8**	**33.5**

Table 4: BLEU scores using different initializations of the XLM embedding layer. XLM is then trained on the respective language pair and used to initialize a UNMT system. Both embeddings are aligned using *VecMap*.

How should static embeddings be integrated in the MLM training? We explore different ways of incorporating the lexical knowledge of pretrained cross-lingual embeddings to the second, masked language modeling stage of our approach (§2.2). Specifically, we keep the aligned embeddings fixed (*frozen*) during XLM training and compare the performance of the final UNMT model to the proposed (*fine-tuned*) method. We point out that, after we transfer the trained MLM to an encoder-decoder model, all layers are trained for UNMT.

Table 4 summarizes our results. The fine-tuning approach, which is adopted in our proposed method, provides a higher performance both in En-Mk and En-Sq, with the improvement being more evident in En-Sq. Our findings generally show that it is preferable to train the bilingual embeddings together with the rest of the model in the MLM step.

6 Related Work

Artetxe et al. (2018c); Lample et al. (2018a) initialize UNMT models with word-by-word transla-

tions, based on a bilingual lexicon inducted in an unsupervised way by the same monolingual data, or simply with cross-lingual embeddings. Lample et al. (2018c) also use pretrained embeddings, learned on joint monolingual corpora of the two languages of interest, to initialize the embedding layer of the encoder-decoder. Lample and Conneau (2019) remove pretrained embeddings from the UNMT pipeline and align language distributions by simply pretraining a MLM on both languages, in order to learn a cross-lingual mapping. However, it has been shown that this pretraining method provides a weak alignment of the language distributions (Ren et al., 2019). While that work identified as a cause the lack of sharing n-gram level cross-lingual information, we address the lack of cross-lingual information at the lexical level.

Moreover, most prior work on UNMT focuses on languages with abundant, high-quality monolingual corpora. In low-resource scenarios though, especially when the languages are not related, pretraining a cross-lingual MLM for unsupervised NMT does not yield good results (Guzmán et al., 2019; Chronopoulou et al., 2020). We propose a method that overcomes this issue by enhancing the MLM with cross-lingual lexical-level representations.

Another line of work tries to enrich the representations of multilingual MLMs with additional knowledge (Wang et al., 2020; Pfeiffer et al., 2020) without harming the already-learned representations. In our work, we identify lexical information as a source of knowledge that is missing from MLMs, especially when it comes to low-resource languages. Surprisingly, static embeddings, such as *fastText*, largely outperform representations extracted by multilingual MLMs in terms of cross-lingual lexical alignment (Vulić et al., 2020). Motivated by this, we aim to narrow the gap between the lexical representations of bilingual MLMs and static embeddings, in order to achieve a higher translation quality, when transferring the MLM to an encoder-decoder UNMT model.

7 Conclusion

We propose a method to improve the lexical ability of a Transformer encoder by initializing its embedding layer with pretrained cross-lingual embeddings. The Transformer is trained for masked language modeling on the language pair of interest. After that, it is used to initialize an encoder/decoder model, which is trained for UNMT and outperforms relevant baselines. Results confirm our intuition that masked language modeling, which provides contextual representations, benefits from cross-lingual embeddings, which capture lexical-level information. In the future, we would like to investigate whether lexical knowledge can be infused to multilingual MLMs. We would also like to experiment with other schemes of training the MLM in terms of how the embedding layer is updated, such as regularizer annealing strategies, which would enable keeping the embeddings relatively fixed, but still allow for some limited training.

8 Ethical Considerations

In this work, we propose a novel unsupervised neural machine translation approach, which is tailored to low-resource languages in terms of monolingual data. We experiment with unsupervised translation between English, Albanian and Macedonian.

For English, we use high-quality data from news articles. The Albanian and Macedonian monolingual data originates from the OSCAR project (Ortiz Suárez et al., 2019). The corpora are shuffled and stripped of all metadata. Therefore, the data should not be easily attributable to specific individuals. Nevertheless, the project offers easy ways to remove data upon request. The En-Sq and En-Mk parallel development and test data are obtained from OPUS (Tiedemann, 2012) and consist of high-quality news articles.

Our work is partly based on training type-level embeddings which are not computationally expensive. However, training cross-lingual masked language models requires significant computational resources. To lower environmental impact, we do not conduct hyper-parameter search and use well-established values for all hyper-parameters.

Acknowledgments

This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement #640550). This work was also supported by DFG (grant FR 2829/4-1). We thank Katerina Margatina, Giorgos Vernikos and Viktor Hangya for their thoughtful comments/suggestions and valuable feedback.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 194–203. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2020. Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2703–2711, Online. Association for Computational Linguistics.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 748–756, Lille, France. PMLR.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The flores evaluation datasets for low-resource machine translation: Nepali–english and sinhala–english. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 6100–6113.

Dan Hendrycks and Kevin Gimpel. 2017. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *ArXiv*.

Philipp Koehn, Marcello Federico, Wade Shen, Nicola Bertoldi, Ondrej Bojar, Chris Callison-Burch, Brooke Cowan, Chris Dyer, Hieu Hoang, Richard Zens, et al. 2006. Open source toolkit for statistical machine translation: Factored translation models and confusion network decoding. In *Final Report of the 2006 JHU Summer Workshop*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, page 7057–7067.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *Workshop on the Challenges in the Management of Large Corpora*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Conference on Machine Translation: Research Papers*, pages 186–191.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 157–163.

Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Explicit cross-lingual pre-training for unsupervised machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 770–779. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 2214–2218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, page 5998–6008.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the International Conference on Machine Learning*, pages 1096–1103.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters.

A Appendix

A.1 Datasets

We remove sentences longer than 100 words after BPE splitting. We split the data using the fastBPE codebase[3].

A.2 Model Configuration

We tie the embedding and output (projection) layers of both LM and NMT models (Press and Wolf, 2017). We use a dropout rate of 0.1 and GELU activations (Hendrycks and Gimpel, 2017). We use the default parameters of Lample and Conneau (2019) in order to train our models.

Regarding the runtimes for the En-Sq experiments: the baseline XLM was trained for 3 days on 8 GPUs while our approach for 6 days and 14h. The experiment with freezing the embeddings provided for faster training, 2 days and 17h. The three methods needed 23h, 21h, and 1d and 8h for the UNMT part, respectively. Fine-tuning with RE-LM took 2 days and 14h on 1 GPU and with our approach it took 1 day and 5h. UNMT for these models took 2 days and 11h, and 13h, respectively. We get a checkpoint every 50K sentences processed by the model.

A.3 Validation Scores of Results

In Table 5 we show the dev scores of the main results, in terms of BLEU scores. This table extends Table 1 of the main paper.

In Table 6, we show the dev scores of the extra fine-tuning experiments we did for the analysis. The table corresponds to Table 4 of the main paper.

	En-Mk		**En-Sq**	
	←	→	←	→
XLM	-	-	30.7	32.0
lexically aligned XLM	**24.6**	**23.3**	**31.9**	**33.8**
RE-LM	25.0	25.7	29.9	32.8
lexically aligned RE-LM	**25.3**	**26.6**	**29.5**	**30.3**

Table 5: UNMT BLEU scores on the development set.

We note that the dev scores are obtained using greedy decoding, while the test scores are obtained with beam search of size 5. We clarify that we train each NMT model using as training criterion the validation BLEU score of the Sq, Mk→En direction, with a patience of 10. We specifically use the `multi-bleu.perl` script from Moses.

[3] https://github.com/glample/fastBPE

Alignment Method	**En-Mk**		**En-Sq**	
	←	→	←	→
lexically aligned MLM				
frozen embeddings	**24.8**	23.0	31.0	32.1
fine-tuned embeddings (ours)	24.6	**23.3**	**31.1**	**32.0**

Table 6: Development BLEU scores using different initializations of the XLM embedding layer.

A.4 *Joint vs VecMap* embeddings.

Using joint embeddings to initialize the MLM, before training it on data from the respective language is less effective for UNMT. This is mostly the case for En-Mk, since the two languages use a different alphabet (Latin and Cyrillic). In this case, simply learning *fastText* embeddings on the concatenation of the two corpora is not useful, because the languages do not have a big lexical overlap.

Alignment Method	**En-Mk**		**En-Sq**	
	←	→	←	→
joint *fastText*	21.5	19.8	32.3	33.1
VECMAP	**25.2**	**22.9**	**32.8**	**33.5**

Table 7: BLEU scores using different initializations of the XLM embedding layer. XLM is then trained on the respective language pair and used to initialize a UNMT system. *Joint fastText embs* refers to jointly learned embeddings following Lample et al. (2018c).