Cross-lingual Language Model Pretraining

Guillaume Lample*
Facebook AI Research
Sorbonne Universités
glample@fb.com

Alexis Conneau*
Facebook AI Research
Université Le Mans
aconneau@fb.com

Abstract

Recent studies have demonstrated the efficiency of generative pretraining for English natural language understanding. In this work, we extend this approach to multiple languages and show the effectiveness of cross-lingual pretraining. We propose two methods to learn cross-lingual language models (XLMs): one unsupervised that only relies on monolingual data, and one supervised that leverages parallel data with a new cross-lingual language model objective. We obtain state-of-the-art results on cross-lingual classification, unsupervised and supervised machine translation. On XNLI, our approach pushes the state of the art by an absolute gain of 4.9% accuracy. On unsupervised machine translation, we obtain 34.3 BLEU on WMT'16 German-English, improving the previous state of the art by more than 9 BLEU. On supervised machine translation, we obtain a new state of the art of 38.5 BLEU on WMT'16 Romanian-English, outperforming the previous best approach by more than 4 BLEU. Our code and pretrained models will be made publicly available.

1 Introduction

Generative pretraining of sentence encoders (Radford et al., 2018; Howard and Ruder, 2018; Devlin et al., 2018) has led to strong improvements on numerous natural language understanding benchmarks (Wang et al., 2018). In this context, a Transformer (Vaswani et al., 2017) language model is learned on a large unsupervised text corpus, and then fine-tuned on natural language understanding (NLU) tasks such as classification (Socher et al., 2013) or natural language inference (Bowman et al., 2015; Williams et al., 2017). Although there has been a surge of interest in learning general-purpose sentence representations, research in that area has been essentially monolingual, and largely focused around English benchmarks (Conneau and Kiela, 2018; Wang et al., 2018). Recent developments in learning and evaluating cross-lingual sentence representations in many languages (Conneau et al., 2018b) aim at mitigating the English-centric bias and suggest that it is possible to build universal cross-lingual encoders that can encode any sentence into a shared embedding space.

In this work, we demonstrate the effectiveness of cross-lingual language model pretraining on multiple cross-lingual understanding (XLU) benchmarks. Precisely, we make the following contributions:

1. We introduce a new unsupervised method for learning cross-lingual representations using cross-lingual language modeling and investigate two monolingual pretraining objectives.

2. We introduce a new supervised learning objective that improves cross-lingual pretraining when parallel data is available.

3. We significantly outperform the previous state of the art on cross-lingual classification, unsupervised machine translation and supervised machine translation.

4. We show that cross-lingual language models can provide significant improvements on the perplexity of low-resource languages.

5. We will make our code and pretrained models publicly available.

*Equal contribution.

2 Related Work

Our work builds on top of Radford et al. (2018); Howard and Ruder (2018); Devlin et al. (2018) who investigate language modeling for pretraining Transformer encoders. Their approaches lead to drastic improvements on several classification tasks from the GLUE benchmark (Wang et al., 2018). Ramachandran et al. (2016) show that language modeling pretraining can also provide significant improvements on machine translation tasks, even for high-resource language pairs such as English-German where there exists a significant amount of parallel data. Concurrent to our work, results on cross-lingual classification using a cross-lingual language modeling approach were showcased on the BERT repository[1]. We compare those results to our approach in Section 5.

Aligning distributions of text representations has a long tradition, starting from word embeddings alignment and the work of Mikolov et al. (2013a) that leverages small dictionaries to align word representations from different languages. A series of follow-up studies show that cross-lingual representations can be used to improve the quality of monolingual representations (Faruqui and Dyer, 2014), that orthogonal transformations are sufficient to align these word distributions (Xing et al., 2015), and that all these techniques can be applied to an arbitrary number of languages (Ammar et al., 2016). Following this line of work, the need for cross-lingual supervision was further reduced (Smith et al., 2017) until it was completely removed (Conneau et al., 2018a). In this work, we take these ideas one step further by aligning distributions of sentences and also reducing the need for parallel data.

There is a large body of work on aligning sentence representations from multiple languages. By using parallel data, Hermann and Blunsom (2014); Conneau et al. (2018b); Eriguchi et al. (2018) investigated zero-shot cross-lingual sentence classification. But the most successful recent approach of cross-lingual encoders is probably the one of Johnson et al. (2017) for multilingual machine translation. They show that a single sequence-to-sequence model can be used to perform machine translation for many language pairs, by using a single shared LSTM encoder and decoder. Their multilingual model outperformed the state of the art on low-resource language pairs, and enabled zero-shot translation. Following this approach, Artetxe and Schwenk (2018) show that the resulting encoder can be used to produce cross-lingual sentence embeddings. Their approach leverages more than 200 million parallel sentences. They obtained a new state of the art on the XNLI cross-lingual classification benchmark (Conneau et al., 2018b) by learning a classifier on top of the fixed sentence representations. While these methods require a significant amount of parallel data, recent work in unsupervised machine translation show that sentence representations can be aligned in a completely unsupervised way (Lample et al., 2018a; Artetxe et al., 2018). For instance, Lample et al. (2018b) obtained 25.2 BLEU on WMT'16 German-English without using parallel sentences. Similar to this work, we show that we can align distributions of sentences in a completely unsupervised way, and that our cross-lingual models can be used for a broad set of natural language understanding tasks, including machine translation.

The most similar work to ours is probably the one of Wada and Iwata (2018), where the authors train a LSTM (Hochreiter and Schmidhuber, 1997) language model with sentences from different languages. They share the LSTM parameters, but use different lookup tables to represent the words in each language. They focus on aligning word representations and show that their approach work well on word translation tasks.

3 Cross-lingual language models

In this section, we present the three language modeling objectives we consider throughout this work. Two of them only require monolingual data (unsupervised), while the third one requires parallel sentences (supervised). We consider N languages. Unless stated otherwise, we suppose that we have N monolingual corpora $\{C_i\}_{i=1...N}$, and we denote by n_i the number of sentences in C_i.

3.1 Shared sub-word vocabulary

In all our experiments we process all languages with the same shared vocabulary created through Byte Pair Encoding (BPE) (Sennrich et al., 2015). As shown in Lample et al. (2018a), this greatly improves the alignment of embedding spaces across languages that share either the same alphabet or anchor tokens such as digits (Smith et al., 2017) or proper nouns. We learn the BPE splits on the concatenation of sentences sampled randomly from

[1] https://github.com/google-research/bert

the monolingual corpora. Sentences are sampled according to a multinomial distribution with probabilities $\{q_i\}_{i=1...N}$, where:

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^N p_j^\alpha} \quad \text{with} \quad p_i = \frac{n_i}{\sum_{k=1}^N n_k}.$$

We consider $\alpha = 0.5$. Sampling with this distribution increases the number of tokens associated to low-resource languages and alleviates the bias towards high-resource languages. In particular, this prevents words of low-resource languages from being split at the character level.

3.2 Causal Language Modeling (CLM)

Our causal language modeling (CLM) task consists of a Transformer language model trained to model the probability of a word given the previous words in a sentence $P(w_t|w_1, \ldots, w_{t-1}, \theta)$. While recurrent neural networks obtain state-of-the-art performance on language modeling benchmarks (Mikolov et al., 2010; Jozefowicz et al., 2016), Transformer models are also very competitive (Dai et al., 2019).

In the case of LSTM language models, back-propagation through time (Werbos, 1990) (BPTT) is performed by providing the LSTM with the last hidden state of the previous iteration. In the case of Transformers, previous hidden states can be passed to the current batch (Al-Rfou et al., 2018) to provide context to the first words in the batch. However, this technique does not scale to the cross-lingual setting, so we just leave the first words in each batch without context for simplicity.

3.3 Masked Language Modeling (MLM)

We also consider the masked language modeling (MLM) objective of Devlin et al. (2018), also known as the Cloze task (Taylor, 1953). Following Devlin et al. (2018), we sample randomly 15% of the BPE tokens from the text streams, replace them by a [MASK] token 80% of the time, by a random token 10% of the time, and we keep them unchanged 10% of the time. Differences between our approach and the MLM of Devlin et al. (2018) include the use of text streams of an arbitrary number of sentences (truncated at 256 tokens) instead of pairs of sentences. To counter the imbalance between rare and frequent tokens (e.g. punctuations or stop words), we also subsample the frequent outputs using an approach similar to Mikolov et al. (2013b): tokens in a text stream are sampled according to a multinomial distribution, whose weights are proportional to the square root of their invert frequencies. Our MLM objective is illustrated in Figure 1.

3.4 Translation Language Modeling (TLM)

Both the CLM and MLM objectives are unsupervised and only require monolingual data. However, these objectives cannot be used to leverage parallel data when it is available. We introduce a new translation language modeling (TLM) objective for improving cross-lingual pretraining. Our TLM objective is an extension of MLM, where instead of considering monolingual text streams, we concatenate parallel sentences as illustrated in Figure 1. We randomly mask words in both the source and target sentences. To predict a word masked in an English sentence, the model can either attend to surrounding English words or to the French translation, encouraging the model to align the English and French representations. In particular, the model can leverage the French context if the English one is not sufficient to infer the masked English words. To facilitate the alignment, we also reset the positions of target sentences.

3.5 Cross-lingual Language Models

In this work, we consider cross-lingual language model pretraining with either CLM, MLM, or MLM used in combination with TLM. For the CLM and MLM objectives, we train the model with batches of 64 streams of continuous sentences composed of 256 tokens. At each iteration, a batch is composed of sentences coming from the same language, which is sampled from the distribution $\{q_i\}_{i=1...N}$ above, with $\alpha = 0.7$. When TLM is used in combination with MLM, we alternate between these two objectives, and sample the language pairs with a similar approach.

4 Cross-lingual language model pretraining

In this section, we explain how cross-lingual language models can be used to obtain:

- a better initialization of sentence encoders for zero-shot cross-lingual classification
- a better initialization of supervised and unsupervised neural machine translation systems
- language models for low-resource languages
- unsupervised cross-lingual word embeddings

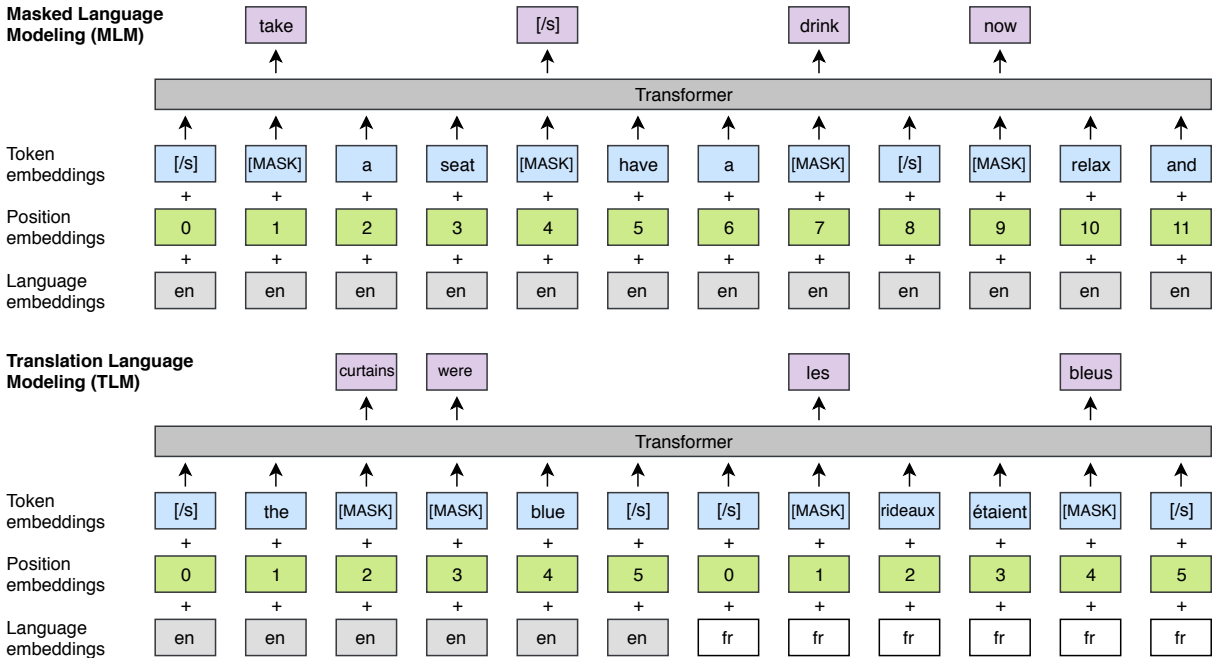

Figure 1: **Cross-lingual language model pretraining.** The MLM objective is similar to the one of Devlin et al. (2018), but with continuous streams of text as opposed to sentence pairs. The TLM objective extends MLM to pairs of parallel sentences. To predict a masked English word, the model can attend to both the English sentence and its French translation, and is encouraged to align English and French representations. Position embeddings of the target sentence are reset to facilitate the alignment.

4.1 Cross-lingual classification

Our pretrained XLM models provide general-purpose cross-lingual text representations. Similar to monolingual language model fine-tuning (Radford et al., 2018; Devlin et al., 2018) on English classification tasks, we fine-tune XLMs on a cross-lingual classification benchmark. We use the cross-lingual natural language inference (XNLI) dataset to evaluate our approach. Precisely, we add a linear classifier on top of the first hidden state of the pretrained Transformer, and fine-tune all parameters on the English NLI training dataset. We then evaluate the capacity of our model to make correct NLI predictions in the 15 XNLI languages. Following Conneau et al. (2018b), we also include machine translation baselines of train and test sets. We report our results in Table 1.

4.2 Unsupervised Machine Translation

Pretraining is a key ingredient of unsupervised neural machine translation (UNMT) (Lample et al., 2018a; Artetxe et al., 2018). Lample et al. (2018b) show that the quality of pretrained cross-lingual word embeddings used to initialize the lookup table has a significant impact on the performance of an unsupervised machine translation model. We propose to take this idea one step further by pretraining the entire encoder and decoder with a cross-lingual language model to bootstrap the iterative process of UNMT. We explore various initialization schemes and evaluate their impact on several standard machine translation benchmarks, including WMT'14 English-French, WMT'16 English-German and WMT'16 English-Romanian. Results are presented in Table 2.

4.3 Supervised Machine Translation

We also investigate the impact of cross-lingual language modeling pretraining for supervised machine translation, and extend the approach of Ramachandran et al. (2016) to multilingual NMT (Johnson et al., 2017). We evaluate the impact of both CLM and MLM pretraining on WMT'16 Romanian-English, and present results in Table 3.

4.4 Low-resource language modeling

For low-resource languages, it is often beneficial to leverage data in similar but higher-resource languages, especially when they share a significant fraction of their vocabularies. For instance, there are about 100k sentences written in Nepali on Wikipedia, and about 6 times more in Hindi. These two languages also have more than 80% of their tokens in common in a shared BPE vocabulary of 100k subword units. We provide in Table 4 a comparison in perplexity between a Nepali lan-

guage model and a cross-lingual language model trained in Nepali but enriched with different combinations of Hindi and English data.

4.5 Unsupervised cross-lingual word embeddings

Conneau et al. (2018a) showed how to perform unsupervised word translation by aligning monolingual word embedding spaces with adversarial training (MUSE). Lample et al. (2018a) showed that using a shared vocabulary between two languages and then applying fastText (Bojanowski et al., 2017) on the concatenation of their monolingual corpora also directly provides high-quality cross-lingual word embeddings (Concat) for languages that share a common alphabet. In this work, we also use a shared vocabulary but our word embeddings are obtained via the lookup table of our cross-lingual language model (XLM). In Section 5, we compare these three approaches on three different metrics: cosine similarity, L2 distance and cross-lingual word similarity.

5 Experiments and results

In this section, we empirically demonstrate the strong impact of cross-lingual language model pretraining on several benchmarks, and compare our approach to the current state of the art.

5.1 Training details

In all experiments, we use a Transformer architecture with 1024 hidden units, 8 heads, GELU activations (Hendrycks and Gimpel, 2016), a dropout rate of 0.1 and learned positional embeddings. We train our models with the Adam optimizer (Kingma and Ba, 2014), a linear warm-up (Vaswani et al., 2017) and learning rates varying from 10^{-4} to 5.10^{-4}.

For the CLM and MLM objectives, we use streams of 256 tokens and a mini-batches of size 64. Unlike Devlin et al. (2018), a sequence in a mini-batch can contain more than two consecutive sentences, as explained in Section 3.2. For the TLM objective, we sample mini-batches of 4000 tokens composed of sentences with similar lengths. We use the averaged perplexity over languages as a stopping criterion for training. For machine translation, we only use 6 layers, and we create mini-batches of 2000 tokens.

When fine-tuning on XNLI, we use mini-batches of size 8 or 16, and we clip the sentence length to 256 words. We use 80k BPE splits and a vocabulary of 95k and train a 12-layer model on the Wikipedias of the XNLI languages. We sample the learning rate of the Adam optimizer with values from 5.10^{-4} to 2.10^{-4}, and use small evaluation epochs of 20000 random samples. We use the first hidden state of the last layer of the transformer as input to the randomly initialized final linear classifier, and fine-tune all parameters. In our experiments, using either max-pooling or mean-pooling over the last layer did not work better than using the first hidden state.

We implement all our models in PyTorch (Paszke et al., 2017), and train them on 64 Volta GPUs for the language modeling tasks, and 8 GPUs for the MT tasks. We use float16 operations to speed up training and to reduce the memory usage of our models.

5.2 Data preprocessing

We use *WikiExtractor*[2] to extract raw sentences from Wikipedia dumps and use them as monolingual data for the CLM and MLM objectives. For the TLM objective, we only use parallel data that involves English, similar to Conneau et al. (2018b). Precisely, we use MultiUN (Ziemski et al., 2016) for French, Spanish, Russian, Arabic and Chinese, and the IIT Bombay corpus (Anoop et al., 2018) for Hindi. We extract the following corpora from the OPUS [3] website Tiedemann (2012): the EUbookshop corpus for German, Greek and Bulgarian, OpenSubtitles 2018 for Turkish, Vietnamese and Thai, Tanzil for both Urdu and Swahili and GlobalVoices for Swahili. For Chinese, Japanese and Thai we use the tokenizer of Chang et al. (2008), the *Kytea*[4] tokenizer, and the *PyThaiNLP*[5] tokenizer respectively. For all other languages, we use the tokenizer provided by Moses (Koehn et al., 2007), falling back on the default English tokenizer when necessary. We use fastBPE[6] to learn BPE codes and split words into subword units. The BPE codes are learned on the concatenation of sentences sampled from all languages, following the method presented in Section 3.1.

[2] https://github.com/attardi/wikiextractor
[3] http://opus.nlpl.eu
[4] http://www.phontron.com/kytea
[5] https://github.com/PyThaiNLP/pythainlp
[6] https://github.com/glample/fastBPE

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. (2018b)	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**

Table 1: **Results on cross-lingual classification accuracy.** Test accuracy on the 15 XNLI languages. We report results for machine translation baselines and zero-shot classification approaches based on cross-lingual sentence encoders. XLM (MLM) corresponds to our unsupervised approach trained only on monolingual corpora, and XLM (MLM+TLM) corresponds to our supervised method that leverages both monolingual and parallel data through the TLM objective. Δ corresponds to the average accuracy.

5.3 Results and analysis

In this section, we demonstrate the effectiveness of cross-lingual language model pretraining. Our approach significantly outperforms the previous state of the art on cross-lingual classification, unsupervised and supervised machine translation.

Cross-lingual classification In Table 1, we evaluate two types of pretrained cross-lingual encoders: an unsupervised cross-lingual language model that uses the MLM objective on monolingual corpora only; and a supervised cross-lingual language model that combines both the MLM and the TLM loss using additional parallel data. Following Conneau et al. (2018b), we include two machine translation baselines: TRANSLATE-TRAIN, where the English MultiNLI training set is machine translated into each XNLI language, and TRANSLATE-TEST where every dev and test set of XNLI is translated to English. We report the XNLI baselines of Conneau et al. (2018b), the multilingual BERT approach of Devlin et al. (2018) and the recent work of Artetxe and Schwenk (2018).

Our fully unsupervised MLM method sets a new state of the art on zero-shot cross-lingual classification and significantly outperforms the supervised approach of Artetxe and Schwenk (2018) which uses 223 million of parallel sentences. Precisely, MLM obtains 71.5% accuracy on average (Δ), while they obtained 70.2% accuracy. By leveraging parallel data through the TLM objective (MLM+TLM), we get a significant boost in performance of 3.6% accuracy, improving even further the state of the art to 75.1%. On the Swahili and Urdu low-resource languages, we outperform the previous state of the art by 6.2% and 6.3% respectively. Using TLM in addition to MLM also improves English accuracy from 83.2% to 85% accuracy, outperforming Artetxe and Schwenk (2018) and Devlin et al. (2018) by 11.1% and 3.6% accuracy respectively.

When fine-tuned on the training set of each XNLI language (TRANSLATE-TRAIN), our supervised model outperforms our zero-shot approach by 1.6%, reaching an absolute state of the art of 76.7% average accuracy. This result demonstrates in particular the consistency of our approach and shows that XLMs can be fine-tuned on any language with strong performance. Similar to the multilingual BERT (Devlin et al., 2018), we observe that TRANSLATE-TRAIN outperforms TRANSLATE-TEST by 2.5% average accuracy, and additionally that our zero-shot approach outperforms TRANSLATE-TEST by 0.9%.

Unsupervised machine translation For the unsupervised machine translation task we consider 3 language pairs: English-French, English-German, and English-Romanian. Our setting is identical to the one of Lample et al. (2018b), except for the initialization step where we use cross-lingual language modeling to pretrain the full model as opposed to only the lookup table.

For both the encoder and the decoder, we consider different possible initializations: CLM pretraining, MLM pretraining, or random initializa-

		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lample et al. (2018b)							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

Table 2: **Results on unsupervised MT.** BLEU scores on WMT'14 English-French, WMT'16 German-English and WMT'16 Romanian-English. For our results, the first two columns indicate the model used to pretrain the encoder and the decoder. " - " means the model was randomly initialized. EMB corresponds to pretraining the lookup table with cross-lingual embeddings, CLM and MLM correspond to pretraining with models trained on the CLM or MLM objectives.

tion, which results in 9 different settings. We then follow Lample et al. (2018b) and train the model with a denoising auto-encoding loss along with an online back-translation loss. Results are reported in Table 2. We compare our approach with the ones of Lample et al. (2018b). For each language pair, we observe significant improvements over the previous state of the art. We re-implemented the NMT approach of Lample et al. (2018b) (EMB), and obtained better results than reported in their paper. We expect that this is due to our multi-GPU implementation which uses significantly larger batches. In German-English, our best model outperforms the previous unsupervised approach by more than 9.1 BLEU, and 13.3 BLEU if we only consider neural unsupervised approaches. Compared to pretraining only the lookup table (EMB), pretraining both the encoder and decoder with MLM leads to consistent significant improvements of up to 7 BLEU on German-English. We also observe that the MLM objective pretraining consistently outperforms the CLM one, going from 30.4 to 33.4 BLEU on English-French, and from 28.0 to 31.8 on Romanian-English. These results are consistent with the ones of Devlin et al. (2018) who observed a better generalization on NLU tasks when training on the MLM objective compared to CLM. We also observe that the encoder is the most important element to pretrain: when compared to pretraining both the encoder and the decoder, pretraining only the decoder leads to a significant drop in performance, while pretraining only the encoder only has a small impact on the final BLEU score.

Pretraining	-	CLM	MLM
Sennrich et al. (2016)	33.9	-	-
ro $\rightarrow$ en	28.4	31.5	35.3
ro $\leftrightarrow$ en	28.5	31.5	35.6
ro $\leftrightarrow$ en + BT	34.4	37.0	**38.5**

Table 3: **Results on supervised MT.** BLEU scores on WMT'16 Romanian-English. The previous state-of-the-art of Sennrich et al. (2016) uses both back-translation and an ensemble model. ro $\leftrightarrow$ en corresponds to models trained on both directions.

Supervised machine translation In Table 3 we report the performance on Romanian-English WMT'16 for different supervised training configurations: mono-directional (ro$\rightarrow$en), bidirectional (ro$\leftrightarrow$en, a multi-NMT model trained on both en$\rightarrow$ro and ro$\rightarrow$en) and bidirectional with back-translation (ro$\leftrightarrow$en + BT). Models with back-translation are trained with the same monolingual data as language models used for pretraining. As in the unsupervised setting, we observe that pretraining provides a significant boost in BLEU score for each configuration, and that pretraining with the MLM objective leads to the best performance. Also, while models with back-translation have access to the same amount of monolingual data as the pretrained models, they are not able to generalize as well on the evaluation sets. Our bidirectional model trained with back-translation obtains the best performance and reaches 38.5 BLEU, outperforming the previous SOTA of Sennrich et al. (2016) (based on back-translation and ensemble models) by more than 4 BLEU.

Low-resource language model In Table 4, we investigate the impact of cross-lingual language modeling for improving the perplexity of a Nepali language model. To do so, we train a Nepali language model on Wikipedia, together with additional data from either English or Hindi. While Nepali and English are distant languages, Nepali and Hindi are similar as they share the same De-

Training languages	Nepali perplexity
Nepali	157.2
Nepali + English	140.1
Nepali + Hindi	115.6
Nepali + English + Hindi	**109.3**

Table 4: **Results on language modeling.** Nepali perplexity when using additional data from a similar language (Hindi) or a distant one (English).

vanagari script and have a common Sanskrit ancestor. When using English data, we reduce the perplexity on the Nepali language model by 17.1 points, going from 157.2 for Nepali-only language modeling to 140.1 when using English. Using additional data from Hindi, we get a much larger perplexity reduction of 41.6. Finally, by leveraging data from both English and Hindi, we reduce the perplexity even more to 109.3 on Nepali. The gains in perplexity from cross-lingual language modeling can be partly explained by the n-grams anchor points that are shared across languages, for instance in Wikipedia articles. The cross-lingual language model can thus transfer the additional context provided by the Hindi or English monolingual corpora through these anchor points to improve the Nepali language model.

Unsupervised cross-lingual word embeddings The MUSE, Concat and XLM (MLM) methods provide unsupervised cross-lingual word embedding spaces that have different properties. In Table 5, we study those three methods using the same word vocabulary and compute the cosine similarity and L2 distance between word translation pairs from the MUSE dictionaries. We also evaluate the quality of the cosine similarity measure via the SemEval'17 cross-lingual word similarity task of Camacho-Collados et al. (2017). We observe that XLM outperforms both MUSE and Concat on cross-lingual word similarity, reaching a Pearson correlation of 0.69. Interestingly, word translation pairs are also far closer in the XLM cross-lingual word embedding space than for MUSE or Concat. Specifically, MUSE obtains 0.38 and 5.13 for cosine similarity and L2 distance while XLM gives 0.55 and 2.64 for the same metrics. Note that XLM embeddings have the particularity of being trained together with a sentence encoder which may enforce this closeness, while MUSE and Concat are based on fastText word embeddings.

	Cosine sim.	L2 dist.	SemEval'17
MUSE	0.38	5.13	0.65
Concat	0.36	4.89	0.52
XLM	**0.55**	**2.64**	**0.69**

Table 5: **Unsupervised cross-lingual word embeddings** Cosine similarity and L2 distance between source words and their translations. Pearson correlation on SemEval'17 cross-lingual word similarity task of Camacho-Collados et al. (2017).

6 Conclusion

In this work, we show for the first time the strong impact of cross-lingual language model (XLM) pretraining. We investigate two unsupervised training objectives that require only monolingual corpora: Causal Language Modeling (CLM) and Masked Language Modeling (MLM). We show that both the CLM and MLM approaches provide strong cross-lingual features that can be used for pretraining models. On unsupervised machine translation, we show that MLM pretraining is extremely effective. We reach a new state of the art of 34.3 BLEU on WMT'16 German-English, outperforming the previous best approach by more than 9 BLEU. Similarly, we obtain strong improvements on supervised machine translation. We reach a new state of the art on WMT'16 Romanian-English of 38.5 BLEU, which corresponds to an improvement of more than 4 BLEU points. We also demonstrate that cross-lingual language model can be used to improve the perplexity of a Nepali language model, and that it provides unsupervised cross-lingual word embeddings. Without using a single parallel sentence, a cross-lingual language model fine-tuned on the XNLI cross-lingual classification benchmark already outperforms the previous supervised state of the art by 1.3% accuracy on average. A key contribution of our work is the translation language modeling (TLM) objective which improves cross-lingual language model pretraining by leveraging parallel data. TLM naturally extends the BERT MLM approach by using batches of parallel sentences instead of consecutive sentences. We obtain a significant gain by using TLM in addition to MLM, and we show that this supervised approach beats the previous state of the art on XNLI by 4.9% accuracy on average. Our code and pretrained models will be made publicly available.

References

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. 2018. Character-level language modeling with deeper self-attention. *arXiv preprint arXiv:1808.04444*.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Kunchukuttan Anoop, Mehta Pratik, and Bhattacharyya Pushpak. 2018. The iit bombay english-hindi parallel corpus. In *LREC*.

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

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

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.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26.

Pi-Chuan Chang, Michel Galley, and Christopher D Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation*, pages 224–232.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *LREC*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv Jegou. 2018a. Word translation without parallel data. In *ICLR*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Zihang Dai, Zhilin Yang, Yiming Yang, William W. Cohen, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Language modeling with longer-term dependency.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *arXiv preprint arXiv:1809.04686*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. *Proceedings of EACL*.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *arXiv preprint arXiv:1606.08415*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 328–339.

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

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

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

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *EMNLP*.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. *NIPS 2017 Autodiff Workshop*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf*.

Prajit Ramachandran, Peter J Liu, and Quoc V Le. 2016. Unsupervised pretraining for sequence to sequence learning. *arXiv preprint arXiv:1611.02683*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *International Conference on Learning Representations*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Wilson L Taylor. 1953. cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Jrg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *LREC*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Takashi Wada and Tomoharu Iwata. 2018. Unsupervised cross-lingual word embedding by multilingual neural language models. *arXiv preprint arXiv:1809.02306*.

Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Paul J Werbos. 1990. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. *Proceedings of NAACL*.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *LREC*.