Cross-lingual Language Model Pretraining

한지윤

INDEX

1 Introduction

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.

3.1 Shared sub-word vocabulary

all languages with the same shared vocabulary

Byte Pair Encoding (BPE)

the alignment of embedding spaces across languages that share either the same alphabet or anchor tokens such as digits or proper nouns

splits on the concatenation of sentences sampled randomly from the monolingual corpora

Sentences are sampled according to a multinomial distribution with probabilities {qi}i=1...N , where:

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

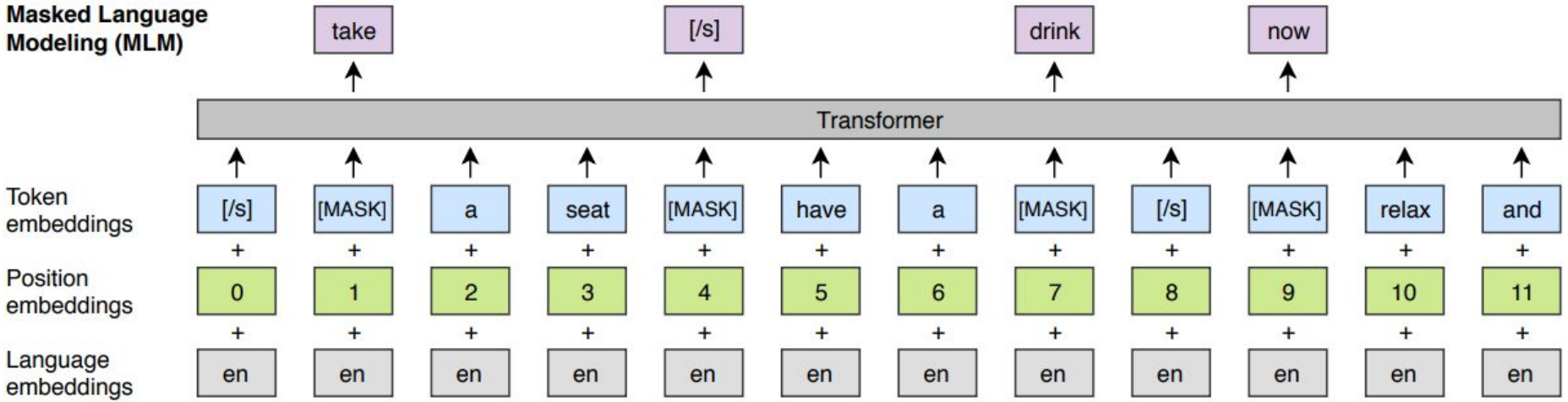
Sampling with this distribution increases the number of tokens associated to low-resource languages and alleviates the bias towards high-resource languages.

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(wt |w1, . . . , wt−1, θ).

3.3 Masked Language Modeling (MLM)

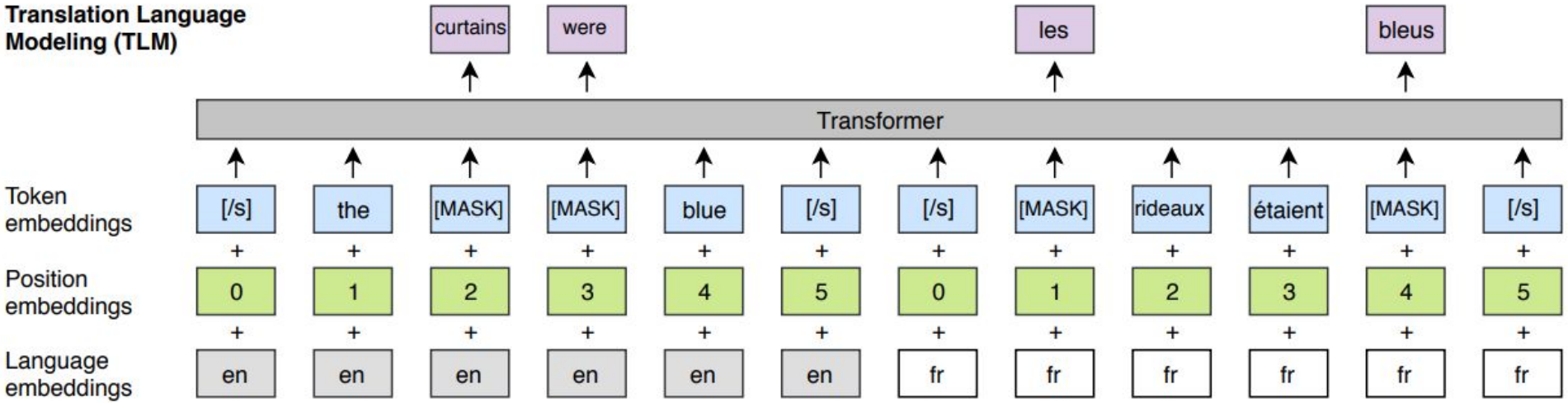
sample randomly 15% of the BPE tokens from the text streams. replace them by a [MASK] token 80% of the time. by a random token 10% of the time. and we keep them unchanged 10% of the time.



Differences : text streams of an arbitrary number of sentences (truncated at 256 tokens) instead of pairs of sentences.

3.4 Translation Language Modeling (TLM)

an extension of MLM, where instead of considering monolingual text streams, we concatenate parallel sentences



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.

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

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

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 crosslingual 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 EnglishRomanian. 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

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 warmup (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 minibatches 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 WikiExtractor2 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 Kytea4 tokenizer, and the PyThaiNLP5 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 fastBPE6 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.

5.3 Results and analysis

- Cross-lingual classification
- Unsupervised machine translation
- Supervised machine translation
- Low-resource language model
- Unsupervised cross-lingual word embeddings

	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.

		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.

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.

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).

	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).

MUSE : Multilingual Unsupervised and Supervised Embeddings
https://github.com/facebookresearch/MUSE

6 Conclusion

MLM pretraining is extremely effective.

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 crosslingual 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.