Cross-lingual Language Model Pre-training

(arXiv2019)

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

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

https://github.com/facebookresearch/XLM

1. Introduction

Cross-lingual Language Model Pre-training

1.1 Two methods to learn cross-lingual language models(XLMs):

- unsupervised（CLM、MLM）
- supervised（TLM）

Cross-lingual Language Model Pre-training

1.2 Contributions:

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

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

Cross-lingual Language Model Pre-training

1.2 Contributions:

- We significantly outperform the previous state of the art on cross-lingual classification, unsupervised machine translation and supervised machine translation.
- We show that cross-lingual language models can provide significant improvements on the perplexity of low-resource languages.

Cross-lingual Language Model Pre-training

1.3 Shared sub-word vocabulary:

- Byte Pair Encoding (BPE) (Sennrich et al., 2015).
- We Sentences are sampled according to a multinomial distribution with probabilities $\{q_i\}_{i=1...N}$, where($\alpha = 0.5$):

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

Cross-lingual Language Model Pre-training

Algorithm 1 Learn BPE operations

```
import re, collections

def get_stats(vocab):
  pairs = collections.defaultdict(int)
  for word, freq in vocab.items():
    symbols = word.split()
    for i in range(len(symbols)-1):
      pairs[symbols[i],symbols[i+1]] += freq
  return pairs

def merge_vocab(pair, v_in):
  v_out = {}
  bigram = re.escape(' '.join(pair))
  p = re.compile(r'(?<!\S)' + bigram + r'(?!\S)')
  for word in v_in:
    w_out = p.sub(''.join(pair), word)
    v_out[w_out] = v_in[word]
  return v_out

vocab = {'l o w </w>' : 5, 'l o w e r </w>' : 2,
         'n e w e s t </w>':6, 'w i d e s t </w>':3}
num_merges = 10
for i in range(num_merges):
  pairs = get_stats(vocab)
  best = max(pairs, key=pairs.get)
  vocab = merge_vocab(best, vocab)
  print(best)
```

For instance :

`aaabdaaabac`

The byte pair "aa" occurs most often, so it will be replaced by a byte that is not used in the data, "Z". Now we have the following data and replacement table:

```
ZabdZabac
Z=aa
```

Then we repeat the process with byte pair "ab", replacing it with Y:

```
ZYdZYac
Y=ab
Z=aa
```

We could stop here, as the only literal byte pair left occurs only once. Or we could continue the process and use recursive byte pair encoding, replacing "ZY" with “X”:

```
XdXac
X=ZY
Y=ab
Z=aa
```

Cross-lingual Language Model Pre-training

2. Cross-lingual language models

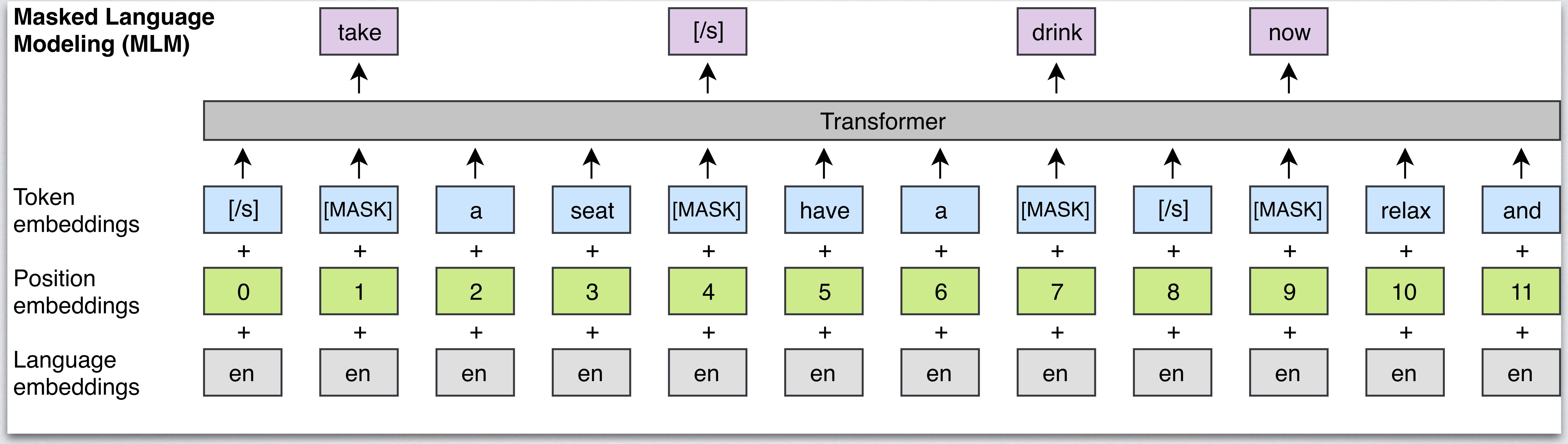
Cross-lingual Language Model Pre-training

2.1 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 $P(w_t | w_1, \ldots, w_{t-1}, \theta)$.

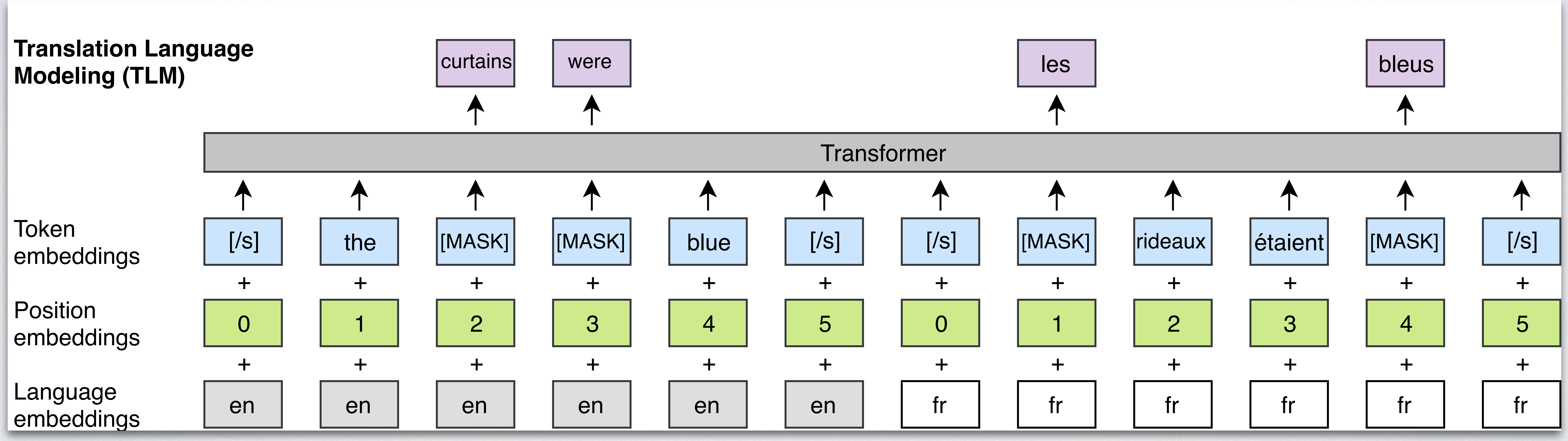
Cross-lingual Language Model Pre-training

2.2 Masked Language Modeling (MLM):



Cross-lingual Language Model Pre-training

2.3 Translation Language Modeling (TLM):



Cross-lingual Language Model Pre-training

2.4 Cross-lingual Language Models:

- **CLM+MLM**(with batches of 64 streams of continuous sentences composed of 256 tokens.sampled from the distribution $\{q_i\}_{i=1...N}$ above, with $\alpha = 0.7$)

- **MLM+TLM**(with a similar approach.)

Cross-lingual Language Model Pre-training

3. XML Pre-training

Cross-lingual Language Model Pre-training

3. Cross-lingual Language Model Pre-training:

- 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

Cross-lingual Language Model Pre-training

3.1 Cross-lingual classification :

- Fine-tune XLMs on a cross-lingual classification benchmark.
- Use the cross-lingual natural language inference (XNLI) dataset to evaluate our approach.
- Evaluate the capacity of our model to make correct NLI predictions in the 15 XNLI languages.
- We also include machine translation baselines of train and test sets. report our results in Table 1.

Cross-lingual Language Model Pre-training

3.2 Unsupervised Machine Translation :

- We propose to take this idea one step further by pre-training 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.
- report our results in Table 2

Cross-lingual Language Model Pre-training

3.3 Supervised Machine Translation :

- We extend the approach of Ramachandran et al. (2016) to multilingual NMT (Johnson et al., 2017).
- report our results in Table 3

Cross-lingual Language Model Pre-training

3.4 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.
- eg. Nepali-Hindi
- report our results in Table 4

Cross-lingual Language Model Pre-training

3.5 Unsupervised cross-lingual word embeddings :

- Conneau et al. (2018a) showed with adversarial training (MUSE)
- Lample et al. (2018a) shared vocabulary between two languages and then applying fastText(Concat)
- report our results(XML) in Table 5

4. Experiments and results

Cross-lingual Language Model Pre-training

4.1 Cross-lingual classification(Table 1) :

	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**

Cross-lingual Language Model Pre-training

4.2 Unsupervised machine translation(Table 2) :

		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**

Cross-lingual Language Model Pre-training

4.3 Supervised machine translation(Table 3) :

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**

Cross-lingual Language Model Pre-training

4.4 Low-resource language model(Table 4) :

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

Cross-lingual Language Model Pre-training

4.5 Unsupervised cross-lingual word embeddings(Table 5) :

	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**

Thanks