Samira Malek

Spring 2023

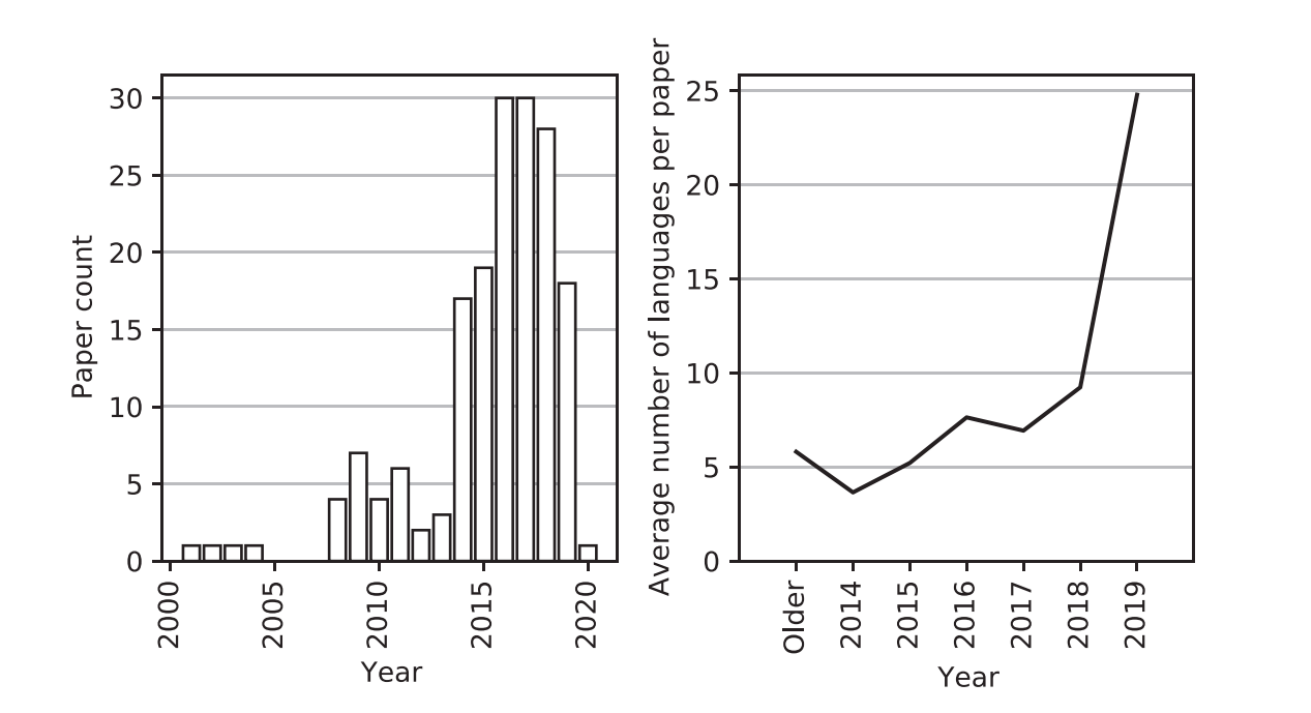
# CROSS-LINGUAL LANGUAGE MODEL PRETRAINING

# WHY MULTILINGUAL MODELS?

- English is the most researched language and it is the single language considered in more than 60% of the papers published.

- Why?
  - Business
  - Help (2010 Haitian earthquake)
  - Fairness

# PROGRESS OF CROSS-LINGUAL LEARNING



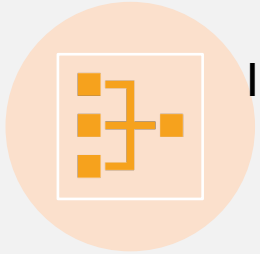

- Fig. Left: Number of surveyed papers per year. Right: Average number of languages used in papers per year[1].

[1]. Pikuliak, Matúš, Marián Šimko, and Mária Bieliková. "Cross-lingual learning for text processing: A survey." *Expert Systems with Applications* 165 (2021): 113765.

# CROSS-LINGUAL LANGUAGE MODELS (XLMS)

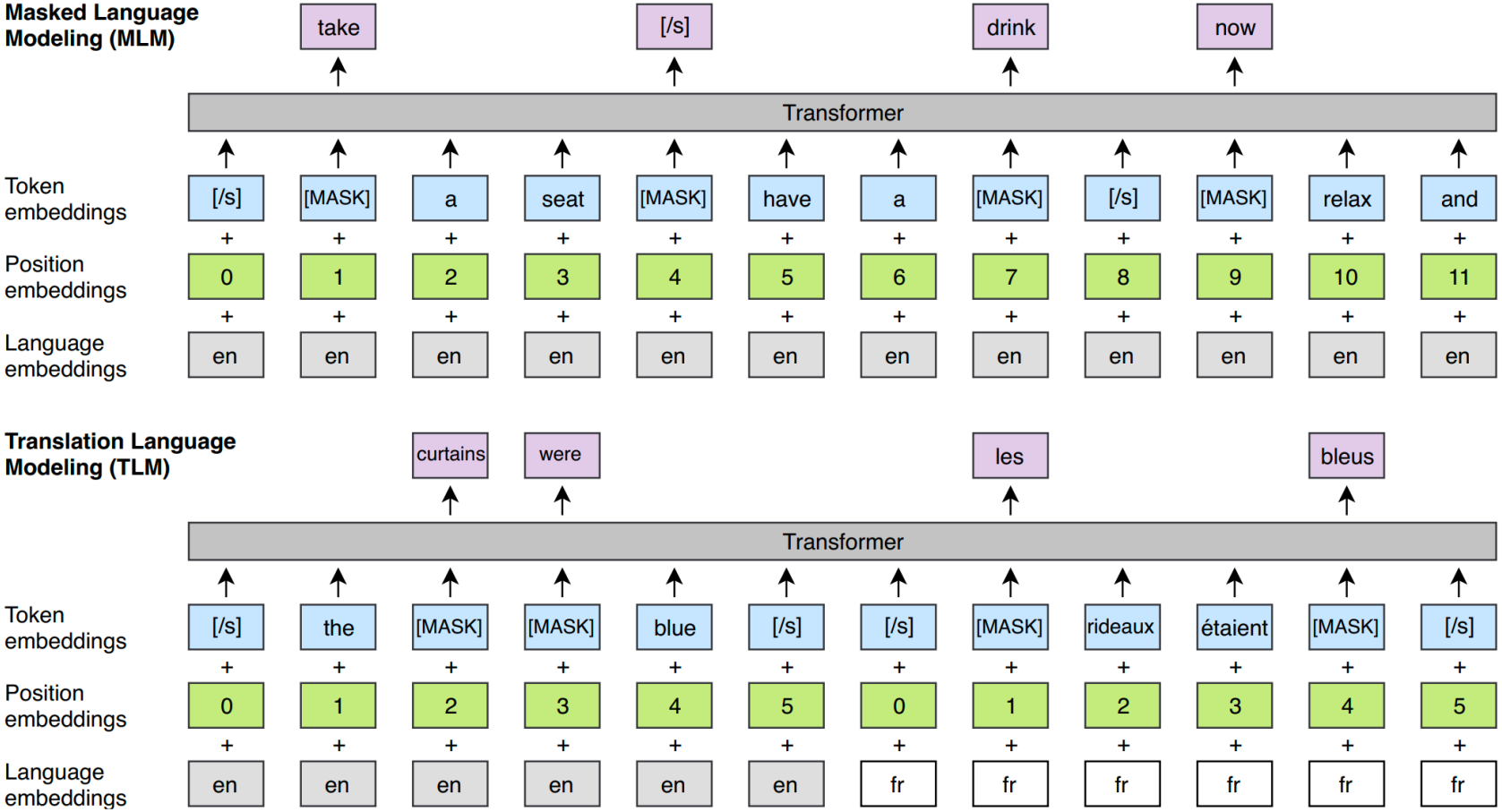
1. Causal Language Modeling (CLM)
2. Masked Language Modeling (MLM)

Unsupervised + Monolingual data



1. Translation Language Modeling (TLM)

Supervised + parallel data


Masked Language Modeling (MLM)
take
[/s]
drink
now
Transformer
Token embeddings
[/s] [MASK] a seat [MASK] have a [MASK] [/s] [MASK] relax and
Position embeddings
0 1 2 3 4 5 6 7 8 9 10 11
Language embeddings
en en en en en en en en en en en en
Translation Language Modeling (TLM)
curtains
were
les
bleus
Transformer
Token embeddings
[/s] the [MASK] [MASK] blue [/s] [/s] [MASK] rideaux étaient [MASK] [/s]
Position embeddings
0 1 2 3 4 5 0 1 2 3 4 5
Language embeddings
en en en en en en fr fr fr fr fr fr

# RESULTS ON CROSS-LINGUAL CLASSIFICATION ACCURACY

| | 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** |

| | | 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** |

Results on unsupervised MT

| 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** |

Results on supervised MT

# LOW-RESOURCE LANGUAGE MODEL

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

# BEYOND XLM: XLM-R

| del | D | #M | #lg | en | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *ne-tune multilingual model on English training set (Cross-lingual Transfer)* | | | | | | | | | | | | | | | | | | | |
| Lample and Conneau (2019) | Wiki+MT | N | 15 | 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 |
| Huang et al. (2019) | Wiki+MT | N | 15 | 85.1 | 79.0 | 79.4 | 77.8 | 77.2 | 77.2 | 76.3 | 72.8 | 73.5 | 76.4 | 73.6 | 76.2 | 69.4 | 69.7 | 66.7 | 75.4 |
| Devlin et al. (2018) | Wiki | N | 102 | 82.1 | 73.8 | 74.3 | 71.1 | 66.4 | 68.9 | 69.0 | 61.6 | 64.9 | 69.5 | 55.8 | 69.3 | 60.0 | 50.4 | 58.0 | 66.3 |
| Lample and Conneau (2019) | Wiki | N | 100 | 83.7 | 76.2 | 76.6 | 73.7 | 72.4 | 73.0 | 72.1 | 68.1 | 68.4 | 72.0 | 68.2 | 71.5 | 64.5 | 58.0 | 62.4 | 71.3 |
| Lample and Conneau (2019) | Wiki | 1 | 100 | 83.2 | 76.7 | 77.7 | 74.0 | 72.7 | 74.1 | 72.7 | 68.7 | 68.6 | 72.9 | 68.9 | 72.5 | 65.6 | 58.2 | 62.4 | 70.7 |
| **XLM-R$_{\text{Base}}$** | CC | 1 | 100 | 85.8 | 79.7 | 80.7 | 78.7 | 77.5 | 79.6 | 78.1 | 74.2 | 73.8 | 76.5 | 74.6 | 76.7 | 72.4 | 66.5 | 68.3 | 76.2 |
| **XLM-R** | CC | 1 | 100 | **89.1** | **84.1** | **85.1** | **83.9** | **82.9** | **84.0** | **81.2** | **79.6** | **79.8** | **80.8** | **78.1** | **80.2** | **76.9** | **73.9** | **73.8** | **80.9** |
| *Translate everything to English and use English-only model (TRANSLATE-TEST)* | | | | | | | | | | | | | | | | | | | |
| BERT-en | Wiki | 1 | 1 | 88.8 | 81.4 | 82.3 | 80.1 | 80.3 | 80.9 | 76.2 | 76.0 | 75.4 | 72.0 | 71.9 | 75.6 | 70.0 | 65.8 | 65.8 | 76.2 |
| RoBERTa | Wiki+CC | 1 | 1 | **91.3** | 82.9 | 84.3 | 81.2 | 81.7 | 83.1 | 78.3 | 76.8 | 76.6 | 74.2 | 74.1 | 77.5 | 70.9 | 66.7 | 66.8 | 77.8 |
| *Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)* | | | | | | | | | | | | | | | | | | | |
| Lample and Conneau (2019) | Wiki | N | 100 | 82.9 | 77.6 | 77.9 | 77.9 | 77.1 | 75.7 | 75.5 | 72.6 | 71.2 | 75.8 | 73.1 | 76.2 | 70.4 | 66.5 | 62.4 | 74.2 |
| *Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)* | | | | | | | | | | | | | | | | | | | |
| Lample and Conneau (2019)† | Wiki+MT | 1 | 15 | 85.0 | 80.8 | 81.3 | 80.3 | 79.1 | 80.9 | 78.3 | 75.6 | 77.6 | 78.5 | 76.0 | 79.5 | 72.9 | 72.8 | 68.5 | 77.8 |
| Huang et al. (2019) | Wiki+MT | 1 | 15 | 85.6 | 81.1 | 82.3 | 80.9 | 79.5 | 81.4 | 79.7 | 76.8 | 78.2 | 77.9 | 77.1 | 80.5 | 73.4 | 73.8 | 69.6 | 78.5 |
| Lample and Conneau (2019) | Wiki | 1 | 100 | 84.5 | 80.1 | 81.3 | 79.3 | 78.6 | 79.4 | 77.5 | 75.2 | 75.6 | 78.3 | 75.7 | 78.3 | 72.1 | 69.2 | 67.7 | 76.9 |
| **XLM-R$_{\text{Base}}$** | CC | 1 | 100 | 85.4 | 81.4 | 82.2 | 80.3 | 80.4 | 81.3 | 79.7 | 78.6 | 77.3 | 79.7 | 77.9 | 80.2 | 76.1 | 73.1 | 73.0 | 79.1 |
| **XLM-R** | CC | 1 | 100 | **89.1** | **85.1** | **86.6** | **85.7** | **85.3** | **85.9** | **83.5** | **83.2** | **83.1** | **83.7** | **81.5** | **83.7** | **81.6** | **78.0** | **78.1** | **83.6** |

[2] Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning at scale." *arXiv preprint arXiv:1911.02116* (2019).

# THANK YOU

# QUESTIONS?