A Brief History of Large Multilingual Language Models

Tianjian Li

Johns Hopkins University

July 12, 2022

1 Chapter 1: Pretrained Language Models

1.1 BERT

In 2018, Jacob Devlin and his team at Google AI published a paper named "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"[4]. BERT has beaten the current best performance on the GLUE[8] benchmark by a very large margin, and has become the backbone structure of language models nowadays.[7][12][13] There are a few reasons that account for the universal effectiveness of BERT, but the two main reasons are:

1. BERT stacks multiple Transformer[2] encoder blocks, which adopts the attention mechanism to capture long range dependencies, allowing each word to dynamically extract syntactic and semantic meaning from every other word. Building on the Transformers, BERT also made a few engineering choices that serves its pretrain objective, including using the sum of the token embedding, segment embedding and positional embedding as the final input representation of a token[1].
2. BERT proposes a pretrain-finetune paradigm, which first trains the model with a unsupervised objective, giving the model parameters better initialization to be finetuned with a supervised objective on downstream tasks with labelled data. Moreover, BERT proposes a **Masked Language Modeling(MLM)** self-supervised(or unsupervised) objective which is effective to teach the model to encode deep bidirectional dependencies.

Note that BERT also proposes a **Next Sentence Prediction(NSP)** unsupervised task, in which the model is given a pair of sentences to detect whether or not the second one is the next sentence of the first one. However subsequent studies[12][13] have confirmed that NSP is not particularly useful, if not detrimental to the performance of downstream tasks, therefore I choose to omit the NSP task when accounting for BERT's effectiveness.

1.2 Masked Language Modeling

Language Modeling aims to calculate the probability density function of a word conditioning on its context. For example, if we are trying to calculate which words is most likely to appear after the context "I ate an sweet apple, it was", a pretrained language model would give us the probability of different words appearing after "I ate an sweet apple, it was". Mathematically speaking, given a context C, a language model aims to compute the probability distribution of the next word X.

$$P(X_i = x|C = (x_1, x_2, ...x_{i-1}))$$

You might notice that here the next word is only allowed to be conditioned on previous words, but what if the word we are trying to model heavily depends on the following words. For example, if we are trying to

[1] A token is similar to a word, only the longer words are divided into sub-words to reduce the amount of vocabulary and learn the meaning of prefixes and suffixes

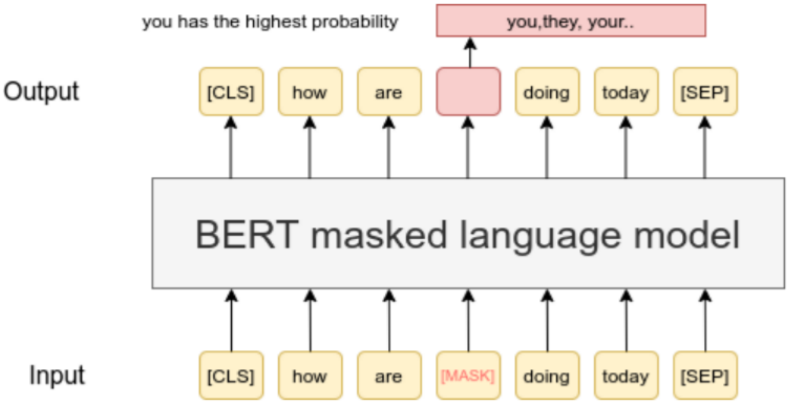


Figure 1: MLM illustration

model the word that goes after ”The following story I am about to tell you is”, we cannot give an accurate prediction of the word if we remain agnostic about the following story. **Masked Language Modeling** aims to solve this issue by replacing the original word with a [MASK] token. While we are modeling the probability distribution of the [MASK] token, we are allowed to condition on bidirectional context:

$$P(X_i = x|C = (x_1, x_2, ...x_{i-1}, x_{i+1}, x_{i+2}, ...x_N))$$

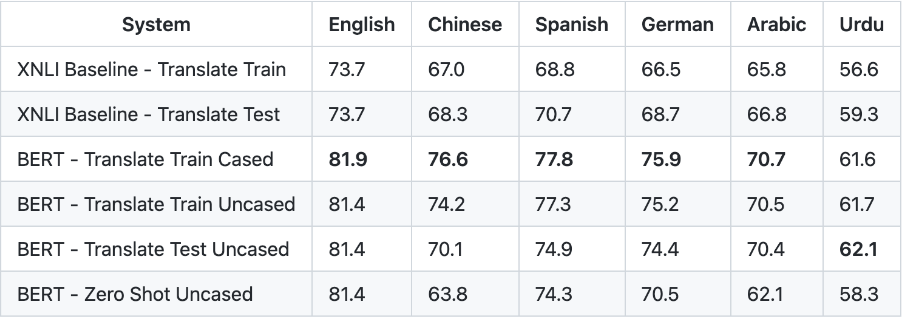
By allowing our language model to learn bidirectionally, BERT has quickly revolutionized the realm of pretraining in NLP, and has cultivated many subsequent work. Models that are built on BERT usually include BERT in its model name. RoBERTa[12] removes the NSP objective and uses dynamic masking, ALBERT[11] aims the reduce the number of parameters, and DeBERTa[14] aims to disentangle the attention for positional embeddings and for word embeddings. However, all of these models are trained on English only data, which limits its usefulness since the majority of person on earth does not speak English. In the next section I will show that BERT not only works on English, but also can be trained on multilingual corpora and possesses an amazing transfer learning ability.

2 Chapter 2: The Curse of Multilinguality

2.1 mBERT and Zero-Shot Transfer

The English-only BERT have beaten the state-of-the-art on the GLUE benchmark by a very large margin, so researchers extended by pretraining BERT on large multilingual corpora, resulting in a model named mBERT(multilingual BERT). Superizingly, language models trained on multilingual corpora have this amazing ability called zero-shot transfer: pretrained models are capable of performing natural language inference(NLI) and natural language understanding(NLU) tasks in different languages by only being fine-tuned on English. Moreover, the zero-shot performance are often comparable to fully supervised methods, albeit still behind a few points. From figure 2 we can see that the performance of mBERT zero-shot(the last row) is comparable to the best Translate-Train(Translating the training data to different languages) performance, which is a supervised method that requires manually or machine translating data. The zero-shot setting is desirable because acquiring high quality data in different languages is often expensive. Therefore we want our model can be taught to generalize from English to other languages and still achieve a high performance. After all, humans excel in generalizing knowledge from one language to another, and we wish to let computers acquire that ability.

Wu and Dredze[9] closely examines the multilinguality of mBERT[4] by extending the downstream task from only XNLI[3] to document classification, named entity recognition part-of-speech tagging and dependency parsing. The main findings of their paper are:



System	English	Chinese	Spanish	German	Arabic	Urdu
XNLI Baseline - Translate Train	73.7	67.0	68.8	66.5	65.8	56.6
XNLI Baseline - Translate Test	73.7	68.3	70.7	68.7	66.8	59.3
BERT - Translate Train Cased	**81.9**	**76.6**	**77.8**	**75.9**	**70.7**	61.6
BERT - Translate Train Uncased	81.4	74.2	77.3	75.2	70.5	61.7
BERT - Translate Test Uncased	81.4	70.1	74.9	74.4	70.4	**62.1**
BERT - Zero Shot Uncased	81.4	63.8	74.3	70.5	62.1	58.3

Figure 2: mBERT performance on XNLI, pasted from mBERT Github repo

1. The zero-shot transfer learning ability of mBERT is universal across all tasks, and outperforms strong zero-shot transfer learning baselines based on LSTM.
2. The knowledge mBERT acquired via pretraining varies across different layers. More importantly, freezing early layers of mBERT leads to better performance to the chosen dowstream tasks.
3. mBERT retains monolingual knowledge despite having strong cross-lingual generalizability.
4. The cross-lingual transfer ability strives under the condition when the source language and the target language shares sub-words. Further studies[15] also confirmed that other multilingual models also benefit from sharing sub-words or even scripts(the same alphabets).

We highlight the zero-shot transfer ability is at its best on NLI and NLU tasks, in which the model is asked to classify sentences, tagging part-of-speech, and extract meaningful sentences from teh original document, as opposed to natural language generation tasks where models are asked to generate words and sentences that fulfill a given requirement. Admittedly, recent studies have been working on enhancing the performance of zero-shot transfer on generation tasks[16][18], it is much harder to let a pretrained multilingual model achieve comparable performance with fully supervised training on zero-shot cross-lingual transfer on generation tasks.

2.2 XLM and XLM-R

language models trained monolingual corpora of different corpora through a self-supervised perspective have shown extraordinary results in providing a better initialization to natural language inference tasks. People began to investigate the potential of pretrained multilingual language models on the canonical area of neural machine translation. At year 2019, researchers at Facebook proposed their novel self-supervised pre-train objective that utilizes parallel corpora in different languages[5].

2.2.1 XLM

XLM[5] proposes a new language model pre-train objective using parallel corpora named **Translation Language Modeling**(TLM). TLM is an extension of MLM, where instead of using monolingual text streams, TLM concatenates parallel sentences and randomly masks in both the source and target sentences. To predict a masked out word in English, the model can attended to the context in English, or the word that have the same semantic meaning in the parallel language, or even the context in the parallel language, encouraging the model to align the representations between the source language and the target language.

Experiment results show that the TLM objective yields major improvements(3.6%) over standard MLM with parallel corpora added advancing the state of the art of the XNLI[3] benchmark to 75.1%. Moreover, obtaining a better alignment not only improves on cross-lingual transfer in classification tasks, but also improves the quality of supervised and unsupervised machine translation.

XLM is trained on data from Wikipedia in the 15 XNLI languages using a MLM objective, and TLM

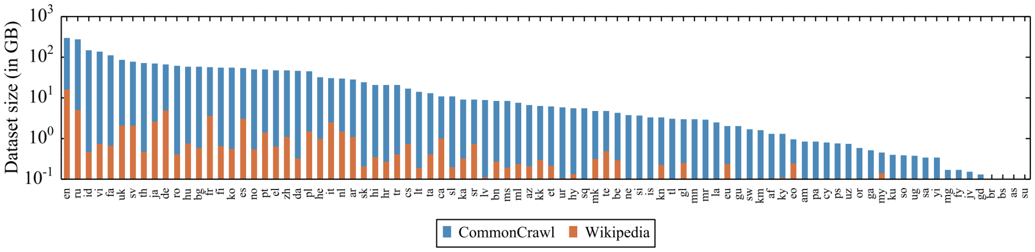


Figure 3: languages and sizes of the language data in the pre-train datasets of XLM-R, pasted from [10]

on additional parallel corpora. We want to highlight that high quality parallel corpora is hard to acquire and even more so for low resource languages.

2.2.2 XLM-R

One of the most recent models release by Facebook(or Meta) is XLM-R[10], which is a multilingual variant of RoBERTa[12]. RoBERTa made a few modifications to BERT, the four main ones are eliminating the need for the NSP task; using dynamic masks that varies across epochs; using longer batches and longer training time; using Byte-Pair Encoding[1] for text tokenizing. The backbone structure of XLM-R is a RoBERTa model. Counter-intuitively, XLM-R eliminates the need for parallel data and only uses unlabeled monolingual corpora extracted from the CommonCrawl dataset containing data in 100 languages.

There are a few key findings of XLM-R, and some of them leads to engineering choices that are crucial to its extraordinary performance. Until now, XLM-R still retains the best performance on the XTREME benchmark(a benchmark evaluating the cross-lingual transferability of language models) compared to other models that has comparable parameters(550 million) to XLM-R.

The first finding of XLM-R is no mystery to language models: **scale matters**. The author finds out that using a larger scale of training corpora leads to a better downstream performance. RoBERTa[12] proposed to train language models for more iterations, which lead to better performance. The XLM-R authors points out that using validation loss(perplexity) as a stopping criterion results in an under-tuned language model. The model's performance on downstream tasks continue to improve even after the validation loss have stopped decreasing. Therefore, researchers that trained XLM-R used a relatively larger model size(L = 24, H = 1024, A = 16, 550M params) and an astonishing amount of 2.5 TB training corpora.

The second finding of XLM-R is what the authors refer to **the curse of multilinguality**[10], which describes the zero-sum game between the number of languages and the performance of the model on low-resourced languages. It is intuitive to assume that given a fixed sized model, more languages means less capacity dedi-

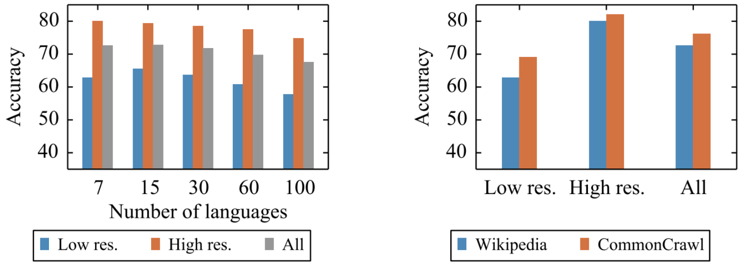


Figure 4: Trade-off between high and low resource language performance. In the right chart, the blue one is trained on wikipedia data containing 100 languages, and the orange one is trained on CC data with only 7 languages. figure pasted from [10]

cated to each language. Since the amount of data for high resource languages far exceeds the amount of data for low resource languages, the performance of low resource languages is severely harmed. Latest research have shown that the curse of multilinguality not only harms the performance of low resource languages, but also negatively affects the performance of high resource languages. From figure 4 we can see that the more the number of pre-training languages, the lower the performance of low resource languages.

We want to highlight that from the left chart in figure 4, the performance of low resource languages increased as we increase the number of pretrained languages increased from 7 to 15. Subsequent studies have found that positive transfer also occurs in that low resource languages are able to transfer knowledge from high resource languages, especially when the two languages share the same script[9][6][15].

The third finding is somewhat a combination of the first two: by scaling the number of parameters and longer training time, we can retain comparable performance of a a multilingual language model to monolingual BERT on the GLUE[8] benchmark, overcoming the curse of multilinguality. We want to highlight that the author uses XLM-7(Model only trained on only 7 languages from CommonCrawl) instead of XLM-R and the baseline is BERT instead of the state-of-the-art RoBERTa, indicating that we can only alleviate the curse of multilinguality but not completely overcome it. Latest studies have also proposed to use language specific modules to over come the curse of multilinguality[17]. Nevertheless, using limited capacity to address unlimited multilingual language semantics will always be a uphill battle.

3 Chapter 3: State of the Art

3.1 MT5

3.2 Our work: MGLM

4 Chapter 4: Going Further

4.1 Parallel Corpora

4.2 Data Augmentation

4.2.1 Intermediate Fine-tuning

4.2.2 Multitask Learning

4.3 Alignment

5 Chapter 5: Epilogue

References

[1] Rico Sennrich, Barry Haddow, and Alexandra Birch. “Neural Machine Translation of Rare Words with Subword Units”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162.

[2] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[3] Alexis Conneau et al. “XNLI: Evaluating Cross-lingual Sentence Representations”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2475–2485. DOI: 10.18653/v1/D18-1269. URL: https://aclanthology.org/D18-1269.

[4] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[5] Guillaume Lample and Alexis Conneau. "Cross-lingual Language Model Pretraining". In: *CoRR* abs/1901.07291 (2019). arXiv: 1901.07291. URL: http://arxiv.org/abs/1901.07291.

[6] Telmo Pires, Eva Schlinger, and Dan Garrette. "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://aclanthology.org/P19-1493.

[7] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019).

[8] Alex Wang et al. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=rJ4km2R5t7.

[9] Shijie Wu and Mark Dredze. "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 833–844. DOI: 10.18653/v1/D19-1077. URL: https://aclanthology.org/D19-1077.

[10] Alexis Conneau et al. "Unsupervised Cross-lingual Representation Learning at Scale". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747.

[11] Zhenzhong Lan et al. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.

[12] Yinhan Liu et al. *Ro{BERT}a: A Robustly Optimized {BERT} Pretraining Approach*. 2020. URL: https://openreview.net/forum?id=SyxS0T4tvS.

[13] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[14] Pengcheng He et al. "{DEBERTA}: {DECODING}-{ENHANCED} {BERT} {WITH} {DISENTANGLED} {ATTENTION}". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[15] Yoshinari Fujinuma, Jordan Boyd-Graber, and Katharina Kann. "Match the Script, Adapt if Multilingual: Analyzing the Effect of Multilingual Pretraining on Cross-lingual Transferability". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1500–1512. DOI: 10.18653/v1/2022.acl-long.106. URL: https://aclanthology.org/2022.acl-long.106.

[16] Kaushal Maurya and Maunendra Desarkar. "Meta-X$_{NLG}$: A Meta-Learning Approach Based on Language Clustering for Zero-Shot Cross-Lingual Transfer and Generation". In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 269–284. DOI: 10.18653/v1/2022.findings-acl.24. URL: https://aclanthology.org/2022.findings-acl.24.

[17] Jonas Pfeiffer et al. "Lifting the Curse of Multilinguality by Pre-training Modular Transformers". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 3479–3495. URL: https://aclanthology.org/2022.naacl-main.255.

[18] Tu Vu et al. *Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation*. 2022. DOI: 10.48550/ARXIV.2205.12647. URL: https://arxiv.org/abs/2205.12647.