Multilinguality in Language Models

Ioannis Partalas

Centrum für Informations- und Sprachverarbeitung
Ludwig-Maximilians-Universität München

May 12, 2023

Seminar: The GPTs and Their Ilk
Prof. Dr. Hinrich Schütze
Haotian Ye, M.Sc.

Table of Contents

1. Motivation

2. Current State
    - Challenges
    - Multilingual Evaluation Results

3. XLM-R (Cross-lingual Language Model from RoBERTa)
    - Architecture
    - Evaluation & Results

4. MAD-X (Multiple ADapters for Cross-lingual Transfer)

5. Literature

Table of Contents

1. Motivation

2. Current State
   - Challenges
   - Multilingual Evaluation Results

3. XLM-R (Cross-lingual Language Model from RoBERTa)
   - Architecture
   - Evaluation & Results

4. MAD-X (Multiple ADapters for Cross-lingual Transfer)

5. Literature

Motivation: Language Models

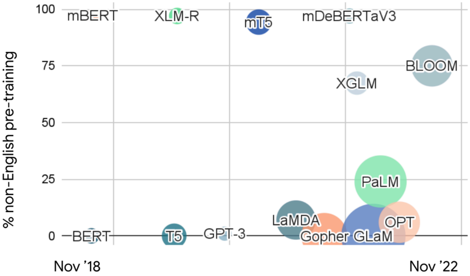


Figure 1: The largest recent models are not becoming significantly more multilingual [Ruder, 2022]

- Cascade of language models
- Language and multi-modal models are demonstrating increasingly powerful capabilities
- Recent models have mostly focused on English and other high-resource languages

Motivation

- The lack of support for non-English languages online can create barriers to access and participation for millions of people, particularly in developing countries
- Crucial to offer information in languages that people understand
- Reinforce the ties between the world and ensure its diversity
- Developing models that work for low-resource languages offsets the existing language segregation

Motivation

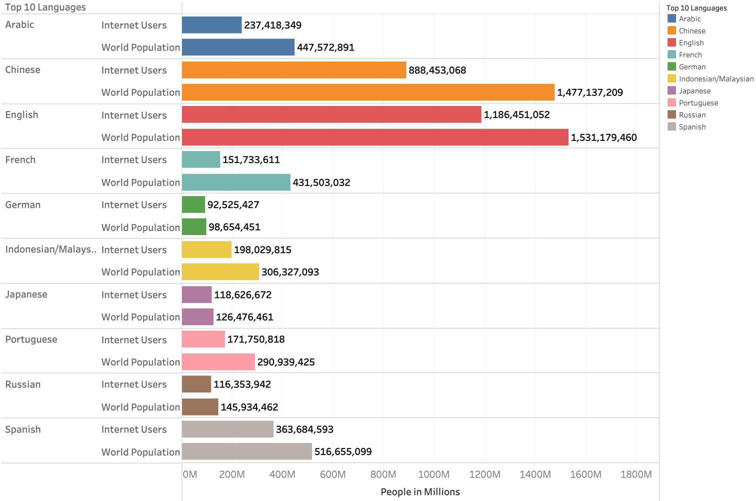


Figure 2: Top ten languages used in the web (figure created based on Internet World Stats [2020])

Motivation

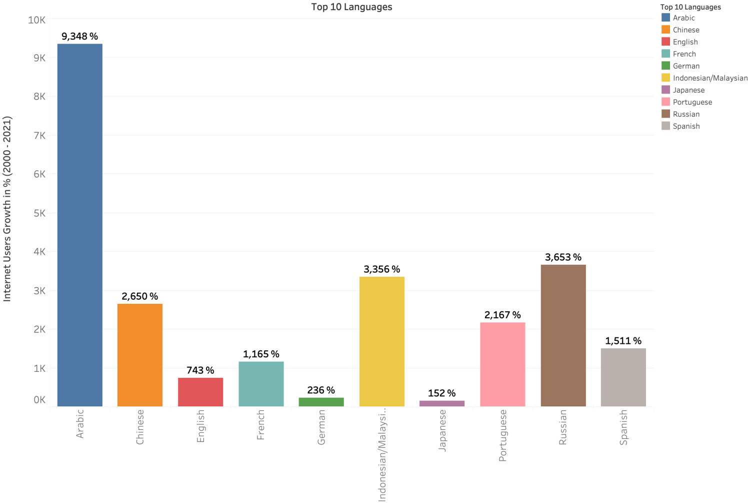


Figure 3: Internet users growth (2000 - 2021) in the web (figure created based on Internet World Stats [2020])

Motivation

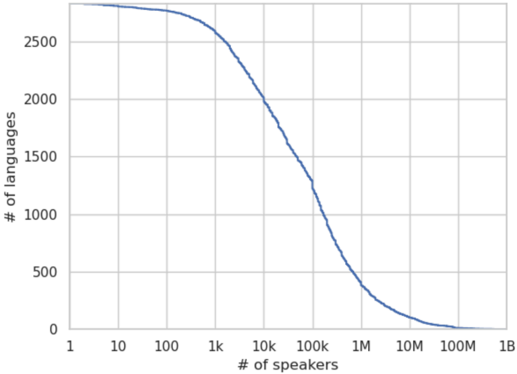


Figure 4: Languages with at least n speakers [van Esch et al., 2022]

- ≈ 7,000 languages spoken in the world
- ≈ 400 languages have more than 1M speakers
- ≈ 1,200 languages have more than 100k speakers

Motivation

- Bender [2011] highlighted the need for language independence in 2011. Reviewing papers published at ACL 2008, she found that 63% of all papers focused on English.
- Ruder et al. [2022] reviewed papers from ACL 2021 and found that almost 70% of papers only evaluate on English
- Many languages in Africa, Asia, and the Americas that are spoken by tens of millions of people have received little research attention [Ruder, 2022]
- Continents (e.g. Africa) with around 2,000 languages or individual countries (e.g. Indonesia) with around 700 languages are incredibly linguistically diverse and at the same time dramatically underserved by current research and technology [Ruder, 2022]

Table of Contents

1 Motivation

2 Current State
- Challenges
- Multilingual Evaluation Results

3 XLM-R (Cross-lingual Language Model from RoBERTa)
- Architecture
- Evaluation & Results

4 MAD-X (Multiple ADapters for Cross-lingual Transfer)

5 Literature

Current State: Challenges

The curse of multilinguality:

- The more languages a model is pre-trained on, the less model capacity is available to learn representations for each language [Conneau et al., 2019]
- Partially improved by increasing the size of a model that allows the model to dedicate more capacity to each language [Goyal et al., 2021]

Current State: Challenges

Lack of pre-training data:

- Languages spoken in Western countries are favored, due to large online existence
- Under-representation of languages with few resources
- According to studies the amount of pre-training data in a language (and its script) correlates with downstream performance for some tasks [Ruder, 2022]

Current State: Challenges

Quality issues in existing multilingual resources:

- Entity names in Wikidata are not in the native script for many under-represented languages while entity spans in WikiAnn are often erroneous [Ruder, 2022]
- Problematic automatically mined resources and automatically aligned corpora for machine translation (e.g. WikiMatrix and CCAligned)

Current State: Multilingual Evaluation Results

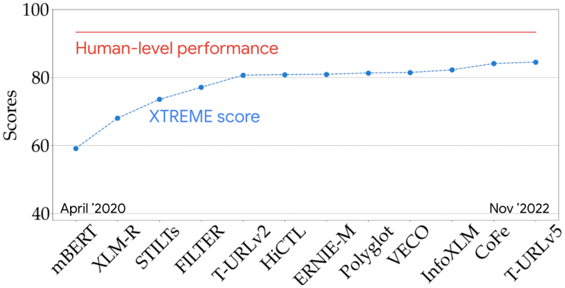


Figure 5: Performance of models on the XTREME leaderboard on 9 tasks and 40 languages [Ruder, 2022]

- Performance has improved steadily and is slowly approaching human-level performance on the benchmark
- Unknown which languages a model was actually evaluated on
- Most tasks and datasets cover few languages

Challenges: Multilingual vs English-centric Models

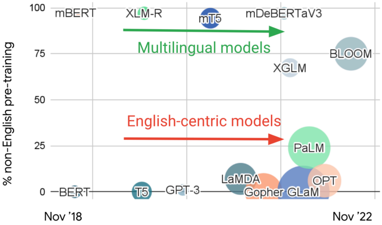


Figure 6: The largest recent models are not becoming significantly more multilingual [Ruder, 2022]

- Two distinct streams of research:
  1. Multilingual models that are trained on multilingual data in many languages
  2. English-centric models that are trained on mostly English data
- English-centric models are increasing in size, still not getting much more multilingual
- Exhibit surprising multilingual capabilities and expected incapabilities
  - GPT-3 and PaLM can translate text between languages with large amounts of data
  - Perform poorly when translating between non-English language pairs or into languages with limited data

Table of Contents

1. Motivation

2. Current State
   - Challenges
   - Multilingual Evaluation Results

3. XLM-R (Cross-lingual Language Model from RoBERTa)
   - Architecture
   - Evaluation & Results

4. MAD-X (Multiple ADapters for Cross-lingual Transfer)

5. Literature

XLM-RoBERTa: Architecture Comparison

Model	**Training objectives**	**Pre-training data**	**Languages**	**Tokenizer & Vocab.**	**Model Size (Params)**	**Misc**
BERT	MLM & NSP	Wikipedia	English	WordPiece & 30K	110M(base) & 335M(large)	Static masking
mBERT	MLM & NSP	Wikipedia	104	WordPiece & 110K	172M	Static masking
XLM	MLM & TLM	Wikipedia & Parallel sentences	100	BPE	?	Language Embeddings, Dynamic masking
RoBERTa	MLM	Wiki, CC-News, OpenWebText, CommonCrawl	English	bBPE & 50K	125M(base) & 355M(large)	Dynamic masking & Larger batch-size & Longer training
XLM-R	MLM	CommonCrawl	100	SPM & 250K	270M(base) & 550M(large)	Dynamic masking & Larger batch-size & Longer training

Table 1: Comparison of key attributes in pre-trained language models

XLM-RoBERTa: MLM

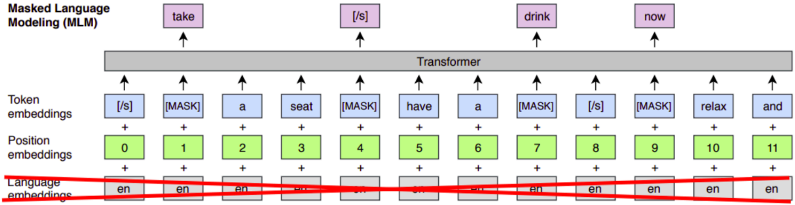


Figure 7: Masked Language Modeling (adapted from Lample and Conneau [2019])

XLM-RoBERTa: Dataset

- CC-100, a clean CommonCrawl Corpus in 100 languages
- Use an internal language identification model in combination with the one from fastText
- Train language models in each language and use it to filter documents
- Significant dataset size increase, especially for low-resource languages

XLM-RoBERTa: Languages

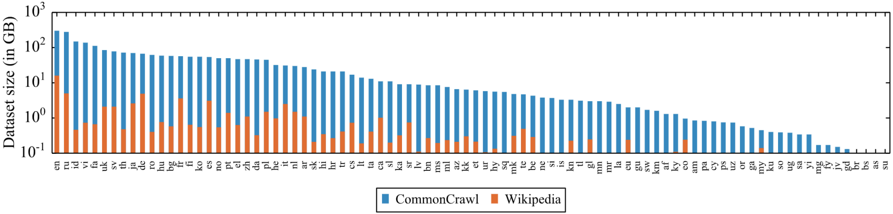


Figure 8: Amount of data in GiB (log-scale) for the 88 languages common in mBERT & XLM-R [Conneau et al., 2019]

- Amount of data increased by several orders of magnitude, in particular for low-resource languages
- Some languages replaced with more commonly used ones such as romanized Hindi and traditional Chinese

XLM-RoBERTa: Evaluation Benchmarks

⇒ 4 evaluation benchmarks

⇒ Cross-lingual understanding: NLI, NER, QA

⇒ Performance of English: Multiple classification tasks

- **XNLI** in 15 languages
- **CoNLL-2002 and CoNLL-2003** in English, Dutch, Spanish and German
- **MLQA** in English, Spanish, German, Arabic, Hindi, Vietnamese and Chinese
- **GLUE** in English on classification tasks such as MNLI & QNLI

XLM-RoBERTa: Ablation Studies

⇒ BERT$_{Base}$ architecture

⇒ Fixed vocabulary of 150K tokens

⇒ 7 languages are fixed (English, French, German, Russian, Chinese, Swahili and Urdu) due to availability of classification and sequence labeling evaluation benchmarks

⇒ Covers a suitable range of language families and includes low-resource languages

⇒ Consider different sets of 7, 15, 30, 60 and all 100 languages

⇒ Conduct most of the analysis on XNLI, representative of findings on other tasks

XLM-RoBERTa: Ablation Studies

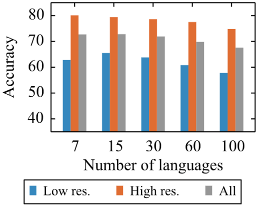


Figure 9: The curse of multilinguality

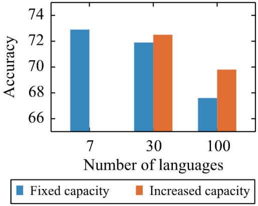


Figure 10: The model capacity effect

- The transfer-interference trade-off
- Low-resource languages benefit from scaling to more languages, until interference kicks in and degrades overall performance

- Adding more languages makes the Transformer wider by increasing the hidden size from 768 to 960 to 1152
- Inadequate added capacity for XLM-100, still lags behind due to higher vocabulary dilution

XLM-RoBERTa: Ablation Studies

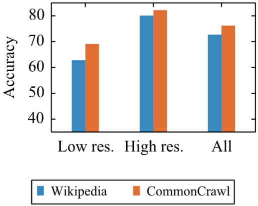


Figure 11: Wikipedia versus CommonCrawl

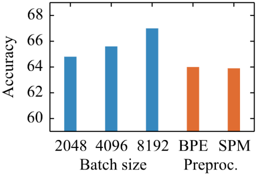


Figure 12: Impact of large-scale training and preprocessing simplification from BPE to SPM

- XLM-7 trained on Wikipedia and CommonCrawl respectively
- Better performance when trained on CC, in particular on low-resource languages

- Increasing batch size and training time enhances model performance
- Tokenization tools used by mBERT and XLM-100 make these models more difficult to use on raw text
- No performance loss for models trained with SPM, compared to models trained with language-specific preprocessing and BPE

XLM-RoBERTa: Ablation Studies

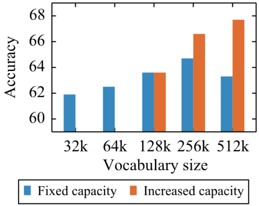


Figure 13: Impact of vocabulary size

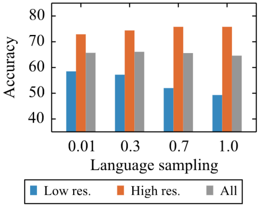
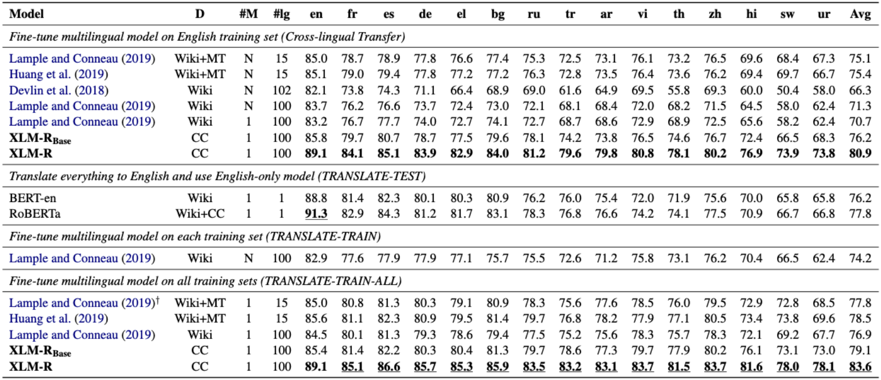


Figure 14: Impact of batch language sampling

- XLM-100 models on Wikipedia with different vocabulary sizes
- Overall number of parameters kept constant by adjusting the width of the transformer

- XLM-100 models trained on Wikipedia with different smoothing factors (α)
- Considering overall performance 0.3 found to be an optimal value for α

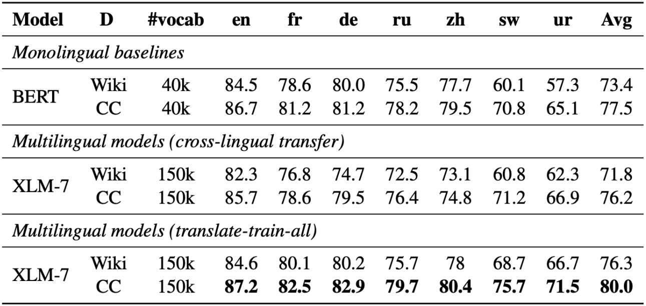
XLM-RoBERTa: Results on XNLI



Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-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**

Figure 15: Results on cross-lingual classification with reports of accuracy on each of the 15 XNLI languages and the average accuracy [Conneau et al., 2019]

XLM-RoBERTa: Monolingual vs Multilingual



Model	D	#vocab	en	fr	de	ru	zh	sw	ur	Avg
Monolingual baselines										
BERT	Wiki	40k	84.5	78.6	80.0	75.5	77.7	60.1	57.3	73.4
	CC	40k	86.7	81.2	81.2	78.2	79.5	70.8	65.1	77.5
Multilingual models (cross-lingual transfer)										
XLM-7	Wiki	150k	82.3	76.8	74.7	72.5	73.1	60.8	62.3	71.8
	CC	150k	85.7	78.6	79.5	76.4	74.8	71.2	66.9	76.2
Multilingual models (translate-train-all)										
XLM-7	Wiki	150k	84.6	80.1	80.2	75.7	78	68.7	66.7	76.3
	CC	150k	**87.2**	**82.5**	**82.9**	**79.7**	**80.4**	**75.7**	**71.5**	**80.0**

Figure 16: Multilingual BERT$_{Based}$ XLM vs monolingual models [Conneau et al., 2019]

- Leveraging training data from multiple languages for a particular task (seemingly) overcomes the capacity dilution problem and leads to better performance

XLM-R: Shortcomings of Fine-tuning Approach and MLMs

Zero-shot transfer to low-resource languages

- Step 1: Train a MLM
- Step 2: Fine-tune a model on a task in a high-resource ***source*** language
- Step 3: Transfer and evaluate on a low-resource ***target*** language

- Training data is expensive and unavailable, especially for low-resource languages
- Multilingual initialisation balances many languages, thus not suited to excel at a specific language at inference time

Language coverage-model capacity trade-off

- XLM-R et al. still perform poorly on cross-lingual transfer across many language pairs, as are unable to represent all languages equally (vocabulary and representation space)
- Scaling up a model to cover all of the world's 7,000+ languages is prohibitive
- Limited capacity is an issue for:
  - high-resource languages - underperform their monolingual variants
  - included low-resource languages & excluded lower-resource languages

Table of Contents

1. Motivation

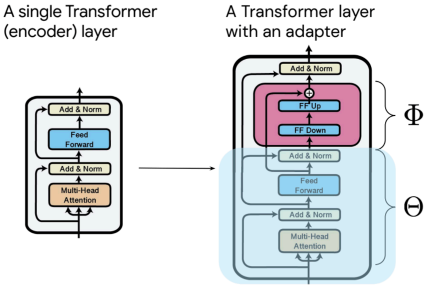
2. Current State
   - Challenges
   - Multilingual Evaluation Results

3. XLM-R (Cross-lingual Language Model from RoBERTa)
   - Architecture
   - Evaluation & Results

4. MAD-X (Multiple ADapters for Cross-lingual Transfer)

5. Literature

MAD-X: Adapters in General



Adapter parameters Φ are **encapsulated** between transformer layers with parameters Θ which are frozen

Figure 17: Transformer layer with an adapter [Ruder, 2022]

- Allocate additional capacity for each language using adapters
- Using a SOTA MLM as foundation, adapt the model to arbitrary tasks and languages by learning modular language- and task-specific representations via adapters
- Small bottleneck layers inserted between a pre-trained model's weights

MAD-X: Adapter Components in MAD-X

Language Adapters

- Trained on unlabelled data of a language using MLM
- Encourages to learn transformations that make the pretrained LM more suitable for a specific language

Invertible Adapters

- Deals with a possible mismatch between the shared multilingual vocabulary and target language vocabulary
- Aim to capture token-level language-specific transformations
- Trained together with the language adapters using MLM on unlabelled data of a specific language
- Parameter-efficient compared to previous approaches

Task Adapters

- Same architecture as language adapters
- Aim to capture knowledge that is task-specific but generalises across languages (language-agnostic)

MAD-X: Training Steps of Adapters in MAD-X

- ***Step 1: Train Language Adapters***
  Train language adapters for the source language and the target language with MLM on Wikipedia

- ***Step 2: Train a Task Adapter***
  Train a task adapter in the source language stacked on top of the source language adapter. The language adapter and the transformer weights are frozen. Only the task adapter is trained

- ***Step 3: Zero-Shot Transfer to Target Language***
  Replace the source language adapter with the target language adapter, while keeping the “language agnostic” task adapter fixed

MAD-X: Architecture of MAD-X

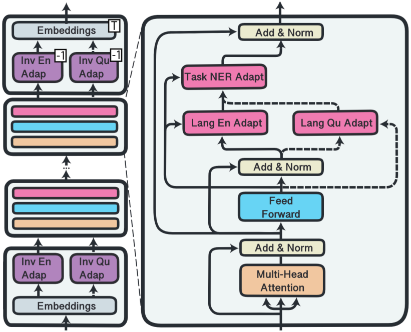


Figure 18: The MAD-X framework inside a Transformer model [Pfeiffer et al., 2020]

Table of Contents

1. Motivation
2. Current State
   - Challenges
   - Multilingual Evaluation Results
3. XLM-R (Cross-lingual Language Model from RoBERTa)
   - Architecture
   - Evaluation & Results
4. MAD-X (Multiple ADapters for Cross-lingual Transfer)
5. Literature

Literature

[1] Emily M Bender. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6, 2011.

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[3] Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.repl4nlp-1.4. URL https://aclanthology.org/2021.repl4nlp-1.4.

[4] Internet World Stats. Internet world users by language. https://www.internetworldstats.com/stats7.htm, 2020. Accessed: 02-05-2023.

[5] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

[6] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*, 2020.

[7] Sebastian Ruder. The State of Multilingual AI. http://ruder.io/state-of-multilingual-ai/, 2022.

[8] Sebastian Ruder, Ivan Vulić, and Anders Søgaard. Square one bias in nlp: Towards a multi-dimensional exploration of the research manifold. *arXiv preprint arXiv:2206.09755*, 2022.

[9] Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera. Writing system and speaker metadata for 2,800+ language varieties. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5035–5046, 2022.