Multilingualism Encourages Recursion

A Transfer Study with mBERT

Andrea de Varda
Roberto Zamparelli

Università di Trento
CIMeC Center for Mind/Brain Sciences
Università degli Studi di Milano Bicocca

Massively Multilingual Models

Massively Multilingual Models (MMMs) are neural networks that can perform a NLP task in multiple languages, relying on a shared set of parameters.

- Transformer-based (XLM, mBERT)
- Zero-shot cross-lingual transfer
  1. Train in N languages $\{L_i\}_{i=1...N}$
  2. Apply to L_{N+1}
- Derived from monolingual models

Massively Multilingual Models

Massively Multilingual Models (MMMs) are neural networks that can perform a NLP task in multiple languages, relying on a shared set of parameters.

- Transformer-based (XLM, mBERT)
- Zero-shot cross-lingual transfer
  1. Train in N languages $\{L_i\}_{i=1...N}$
  2. Apply to L_{N+1}
- Derived from monolingual models

Massively Multilingual Models

Massively Multilingual Models (MMMs) are neural networks that can perform a NLP task in multiple languages, relying on a shared set of parameters.

- Transformer-based (XLM, mBERT)
- Zero-shot cross-lingual transfer
  1. Train in N languages $\{L_i\}_{i=1...N}$
  2. Apply to L_{N+1}
- Derived from monolingual models

BERT

BERT is a 12-layer language representation model designed to produce deep, bidirectional representations from unlabeled text relying on both left and right context across layers.

- Masked language modeling objective (~ cloze procedure)

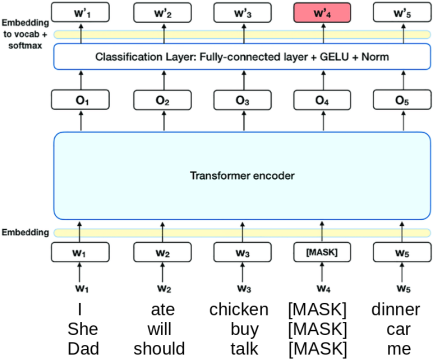


Figure: Monolingual BERT

[Image adapted from https://towardsdatascience.com]

Multilingual BERT

BERT is a 12-layer language representation model designed to produce deep, bidirectional representations from unlabeled text relying on both left and right context across layers.

- Masked language modeling objective (~ cloze procedure)

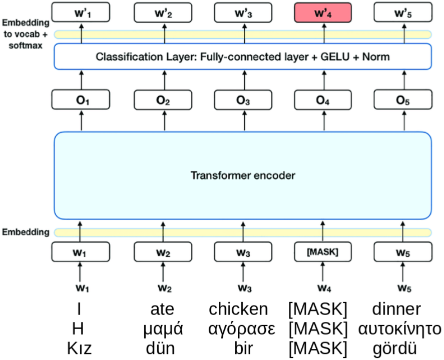
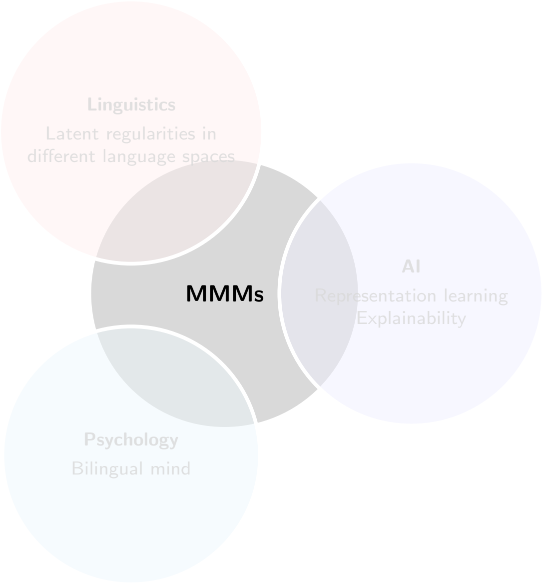
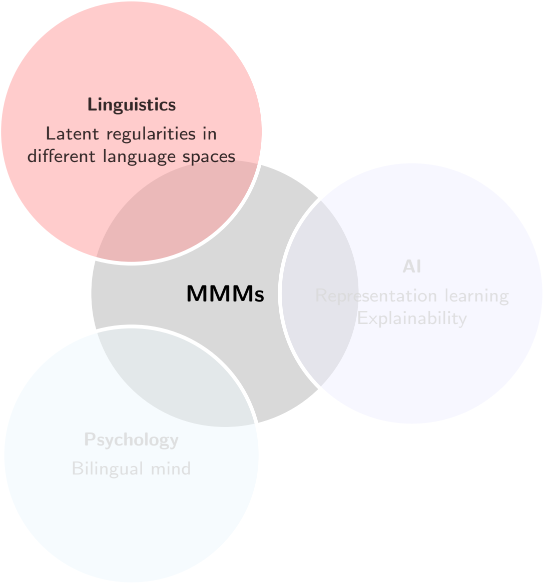


Figure: Multilingual BERT

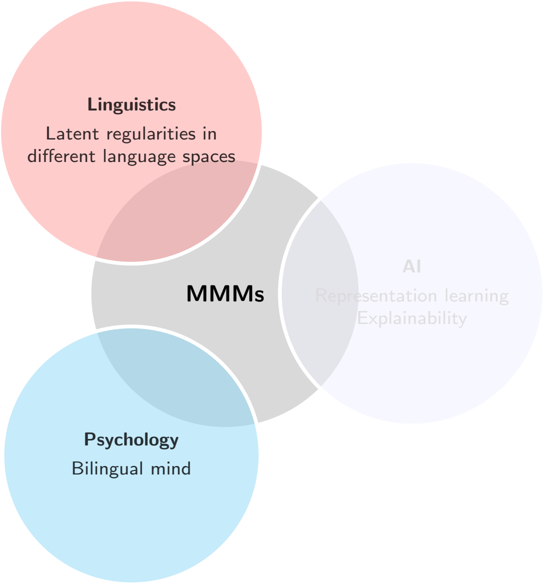
Relevance in NLP and Cognitive Sciences



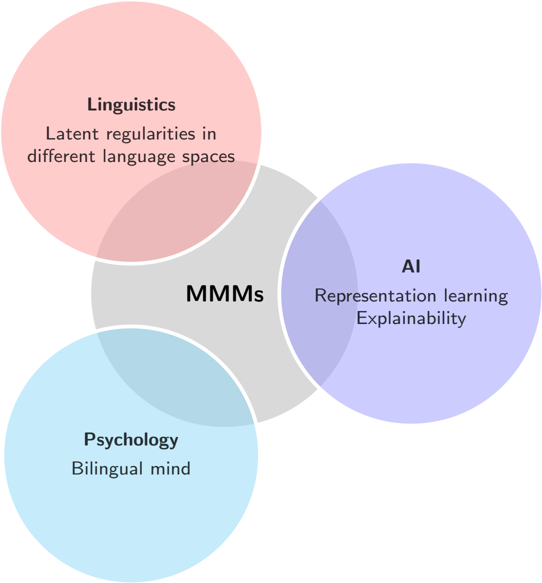
Relevance in NLP and Cognitive Sciences



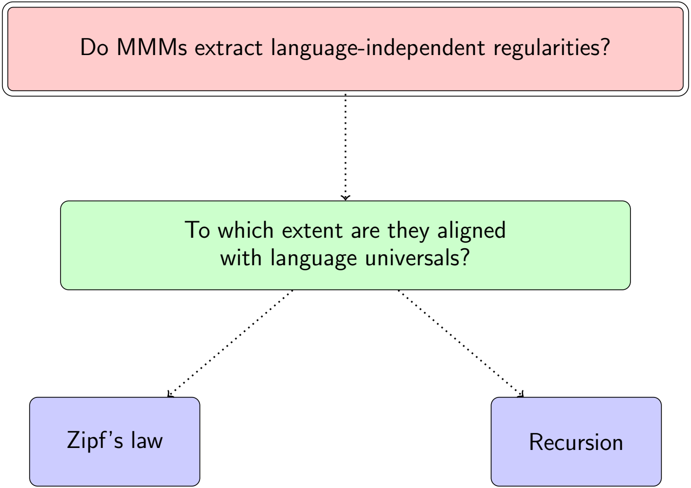
Relevance in NLP and Cognitive Sciences



Relevance in NLP and Cognitive Sciences



Research question



Zipf's Law

The words' frequency of occurrence is inversely related to their frequency rank.

- $Freq(w) \propto \frac{C}{rank(w)}$

Not language-specific:

- Mathematical formulas (Greiner-Petter et al., 2020)
- Musical notation (Zanette, 2006)
- Social sciences (Pustet, 2004)

Recursion

- Biology-driven cross-linguistic universal
- Narrow faculty of language ⇔ recursion (Hauser et al., 2002)
  - Uniquely human component of natural languages
- Allows languages to be generative and productive
- Recursive rules can be reapplied to their own output

Rule: S → a S b

S
a S b
a a S b b
a a a S b b b
a a a a S b b b b
a a a a a S b b b b b
a a a a a a S b b b b b b

Recursion

- Biology-driven cross-linguistic universal
- Narrow faculty of language ⇔ recursion (Hauser et al., 2002)
  - Uniquely human component of natural languages
- Allows languages to be generative and productive
- Recursive rules can be reapplied to their own output

Rule: `S → a S b`

```
          S
        a S b
      a a S b b
    a a a S b b b
  a a a a S b b b b
a a a a a S b b b b b
a a a a a a S b b b b b b
```

Recursion

- Biology-driven cross-linguistic universal
- Narrow faculty of language ⇔ recursion (Hauser et al., 2002)
  - Uniquely human component of natural languages
- Allows languages to be generative and productive
- Recursive rules can be reapplied to their own output

Rule: S → a S b

S
a S b
a a S b b
a a a S b b b
a a a a S b b b b
a a a a a S b b b b b
a a a a a a S b b b b b b

Recursion

- Biology-driven cross-linguistic universal
- Narrow faculty of language ⇔ recursion (Hauser et al., 2002)
  - Uniquely human component of natural languages
- Allows languages to be generative and productive
- Recursive rules can be reapplied to their own output

Rule: `S → a S b`

```
            S
          a S b
        a a S b b
      a a a S b b b
    a a a a S b b b b
  a a a a a S b b b b b
a a a a a a S b b b b b b
```

Recursion

- Biology-driven cross-linguistic universal
- Narrow faculty of language ⇔ recursion (Hauser et al., 2002)
  - Uniquely human component of natural languages
- Allows languages to be generative and productive
- Recursive rules can be reapplied to their own output

Rule: S → a S b

```
                  S
                a S b
              a a S b b
            a a a S b b b
          a a a a S b b b b
        a a a a a S b b b b b
      a a a a a a S b b b b b b
```

Recursion

- Biology-driven cross-linguistic universal
- Narrow faculty of language ⇔ recursion (Hauser et al., 2002)
  - Uniquely human component of natural languages
- Allows languages to be generative and productive
- Recursive rules can be reapplied to their own output

Rule: S → a S b

```
            S
          a S b
        a a S b b
      a a a S b b b
    a a a a S b b b b
  a a a a a S b b b b b
a a a a a a S b b b b b b
```

Recursion

- Biology-driven cross-linguistic universal
- Narrow faculty of language ⇔ recursion (Hauser et al., 2002)
  - Uniquely human component of natural languages
- Allows languages to be generative and productive
- Recursive rules can be reapplied to their own output

Rule: S → a S b

```
            S
          a S b
        a a S b b
      a a a S b b b
    a a a a S b b b b
  a a a a a S b b b b b
a a a a a a S b b b b b b
```

Recursion

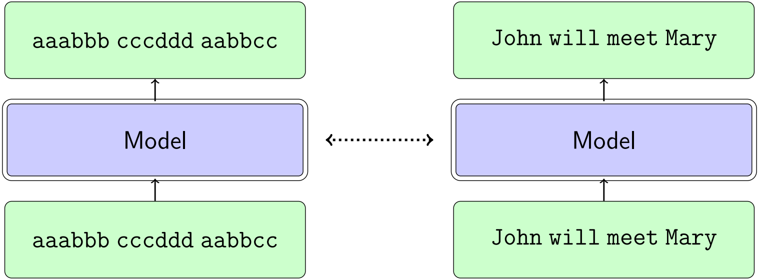
- Biology-driven cross-linguistic universal
- Narrow faculty of language ⇔ recursion (Hauser et al., 2002)
  - Uniquely human component of natural languages
- Allows languages to be generative and productive
- Recursive rules can be reapplied to their own output

Rule: `S → a S b`

```
            S
          a S b
        a a S b b
      a a a S b b b
    a a a a S b b b b
  a a a a a S b b b b b
a a a a a a S b b b b b b
```

Transfer learning to uncover relational structures

Transfer learning has been proposed as a tool for analyzing the encoding of grammatical structures in neural language models (Papadimitriou and Jurafsky, 2020).



Procedure

Adaptation of BERT's native MLM functionality.

- MLM: 15% of the tokens are masked
- → all the tokens are masked iteratively in evaluation mode
    - Comprehensive sentence-wise metric

t_1	[MASK] should buy a car	[P1]
t_2	John [MASK] buy a car	[P1, P2]
t_3	John should [MASK] a car	[P1, P2, P3]
t_4	John should buy [MASK] car	[P1, P2, P3, P4]
t_5	John should buy a [MASK]	[P1, P2, P3, P4, P5]

Figure: Unfolding of the iterative token-level cloze task for every timestep t in a sample sequence.

Procedure

Adaptation of BERT's native MLM functionality.

- MLM: 15% of the tokens are masked
- → all the tokens are masked iteratively in evaluation mode
  - Comprehensive sentence-wise metric

t_1	[MASK] should buy a car	[P1]
t_2	John [MASK] buy a car	[P1, P2]
t_3	John should [MASK] a car	[P1, P2, P3]
t_4	John should buy [MASK] car	[P1, P2, P3, P4]
t_5	John should buy a [MASK]	[P1, P2, P3, P4, P5]

Figure: Unfolding of the iterative token-level cloze task for every timestep t in a sample sequence.

Procedure

Adaptation of BERT's native MLM functionality.

- MLM: 15% of the tokens are masked
- → all the tokens are masked iteratively in evaluation mode
  - Comprehensive sentence-wise metric

t_1	`[MASK] should buy a car`	`[P1]`
t_2	`John [MASK] buy a car`	`[P1, P2]`
t_3	`John should [MASK] a car`	`[P1, P2, P3]`
t_4	`John should buy [MASK] car`	`[P1, P2, P3, P4]`
t_5	`John should buy a [MASK]`	`[P1, P2, P3, P4, P5]`

Figure: Unfolding of the iterative token-level cloze task for every timestep t in a sample sequence.

Procedure

Adaptation of BERT's native MLM functionality.

- MLM: 15% of the tokens are masked
- → all the tokens are masked iteratively in evaluation mode
  - Comprehensive sentence-wise metric

t_1	`[MASK] should buy a car`	`[P1]`
t_2	`John [MASK] buy a car`	`[P1, P2]`
t_3	`John should [MASK] a car`	`[P1, P2, P3]`
t_4	`John should buy [MASK] car`	`[P1, P2, P3, P4]`
t_5	`John should buy a [MASK]`	`[P1, P2, P3, P4, P5]`

Figure: Unfolding of the iterative token-level cloze task for every timestep t in a sample sequence.

Procedure

Adaptation of BERT's native MLM functionality.

- MLM: 15% of the tokens are masked
- → all the tokens are masked iteratively in evaluation mode
  - Comprehensive sentence-wise metric

t_1	`[MASK] should buy a car`	`[P1]`
t_2	`John [MASK] buy a car`	`[P1, P2]`
t_3	`John should [MASK] a car`	`[P1, P2, P3]`
t_4	`John should buy [MASK] car`	`[P1, P2, P3, P4]`
t_5	`John should buy a [MASK]`	`[P1, P2, P3, P4, P5]`

Figure: Unfolding of the iterative token-level cloze task for every timestep t in a sample sequence.

Experiment

We evaluated the predictive behaviour of BERT and mBERT on a set of four artificial corpora (1,000 sequences each) with increasing structural complexity.

Nested brackets

Paired, hierarchically nested tokens

Flat brackets

Paired, non-nested tokens

Zipf corpus

Sampled from Zipf's distribution

Random corpus

Sampled from uniform distribution

Nested brackets

Sequences of nested matching symbols.

- Probabilistic Context-Free Grammar (PCFG)
  - Nested dependencies
  - Pairing arcs never cross

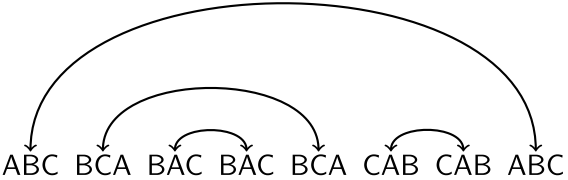


Figure: Example of a NESTED BRACKETS sequence.

Flat Brackets

Non-nested dependency pairing.

- Random shuffling
  - Dependencies do not necessarily nest
  - Pairing arcs may cross
- Same tokens, same length

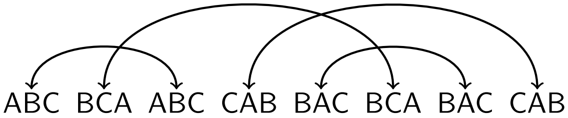


Figure: Example of a FLAT BRACKETS sequence.

Zipfian and random corpora

Zipfian corpus

Tokens sampled from a Zipfian distribution.

Random corpus

Tokens sampled from a uniform distribution.

Hierarchy of structuredness

NESTED BRACKETS > FLAT BRACKETS > ZIPFIAN > RANDOM

Corpus	**Zipf**	**Pairing**	**Nesting**
Nested brackets	✓	✓	✓
Flat brackets	✓	✓	
Zipf's corpus	✓		
Random corpus			

Table: Featural summary of the structural and mathematical properties of the four corpora.

Results

mBERT

NESTED BRACKETS > FLAT BRACKETS > ZIPFIAN > RANDOM

BERT

FLAT BRACKETS > NESTED BRACKETS > ZIPFIAN > RANDOM

	Model			
	BERT		mBERT	
Corpus	Mean	SD	Mean	SD
Random corpus	0.0121	0.0137	0.0094	0.0135
Zipf's corpus	0.0253	0.0457	0.0250	0.0525
Flat brackets	**0.6784**	0.1780	0.6353	0.1558
Nested brackets	0.6576	0.1677	**0.6417**	0.1536

Table: Results of the transfer.

Discussion

Zipf > random

Non-uniform distribution of the tokens

Flat > Zipf

Pairwise head-dependency type structures

Nested > Flat (mBERT)

Hierarchical and recursive representations

Flat > Nested (BERT)

~~Hierarchical and recursive representations~~

Conclusion

Analysis of MMMs' generalizations by quantifying their alignment to different cornerstones in quantitative and theoretical linguistics.

mBERT

- Zipfian distribution of tokens
- Head-dependency type structures
- Hierarchy and recursion

BERT

- Zipfian distribution of tokens
- Head-dependency type structures
- ~~Hierarchy and recursion~~

Deep learning and theoretical linguistics

Some of the representations learnt by multilingual deep learning models are not very distant from the ones that were discovered by theoretical linguists, highlighting the importance of a deeper cross-disciplinary integration between these two fields.

Limitations & future directions

- "Behavioural" results chould be complemented studying the model's inner workings
- Repetition is not a natural way to mark constituent boundaries
- Test on **natural language**

  ABC BCA BAC BAC BCA ABC $\neq$ *The fox that chased the dog fell*

Thank you for your attention!