Polarity-sensitivity in pre-trained language models

Lisa Bylinina, *Bookarang*
(joint work with Alexey Tikhonov, *Yandex*)

ILFC Seminar
14 December 2021

Negative polarity items (NPIs): Intro

Mary didn't buy **any** books.

*Mary bought **any** books.

Negative polarity items (NPIs): Intro

Mary didn't buy **any** books.

*Mary bought **any** books.

No boxes contain **any** plates.

*Some boxes contain **any** plates.

Negative polarity items (NPIs): Intro

Mary didn't buy **any** books.
*Mary bought **any** books.

No boxes contain **any** plates.
*Some boxes contain **any** plates.

Few people had **any** thoughts.
*Many people had **any** thoughts.

Negative polarity items (NPIs): Intro

Mary didn't buy **any** books.
*Mary bought **any** books.

No boxes contain **any** plates.
*Some boxes contain **any** plates.

Few people had **any** thoughts.
*Many people had **any** thoughts.

The use of NPIs is restricted to **negative contexts**.

Negative polarity items (NPIs): Intro

Mary didn't buy **any** books.
*Mary bought **any** books.

No boxes contain **any** plates.
*Some boxes contain **any** plates.

Few people had **any** thoughts.
*Many people had **any** thoughts.

The use of NPIs is restricted to **negative contexts**.
What makes a context negative?

Defining ‘negative context’...

- **Lexically?**
  w.r.t. ‘negative words’ like *not*, *no*, *few*
- **Syntactically?**
  w.r.t. certain types of ‘negative’ projections
- **Semantically?**
  w.r.t. a meaning aspect of the context
  - its monotonicity profile
    (since Fauconnier 1975, 1978; Ladusaw 1979)
  - some other meaning aspect:
    anti-additivity, non-veridicality etc.;
    (Zwarts 1996, Giannakidou 1999 a.o.)

Monotonicity: a 1-minute intro

No students <u>cook</u>

$$\textsc{no}(\textsc{student})(\underline{\textsc{cook}}): \textsc{set}(\textsc{student}) \cap \textsc{set}(\textsc{cook}) = \emptyset$$

Some students <u>cook</u>

$$\textsc{some}(\textsc{student})(\underline{\textsc{cook}}): \textsc{set}(\textsc{student}) \cap \textsc{set}(\textsc{cook}) \neq \emptyset$$

Monotonicity: a 1-minute intro

No students <u>cook</u>

$\textsc{no}(\textsc{student})(\underline{\textsc{cook}})$: $\textsc{set}(\textsc{student}) \cap \textsc{set}(\textsc{cook}) = \emptyset$

Some students <u>cook</u>

$\textsc{some}(\textsc{student})(\underline{\textsc{cook}})$: $\textsc{set}(\textsc{student}) \cap \textsc{set}(\textsc{cook}) \neq \emptyset$

No students cook $\rightarrow$ No students cook rice

Some students cook $\leftarrow$ Some students cook rice

Exactly 3 students cook ? Exactly 3 students cook rice

Monotonicity and NPIs: what we know

- Human judgments about monotonicity and on NPI acceptability are graded (Geurts 2003; Sanford et al. 2007; Chemla et al. 2011; McNabb et al. 2016; Denić et al. 2021)
  - scope of *no* perceived as DE 72% of the time
  - *at most* – 56% of the time
  - *less than* and *at most* differ by 11%
- Individual's judgments of monotonicity are good predictors of their judgments of NPI grammaticality (Chemla et al. 2011)
- The presence of an NPI affects judgments of monotonicity (Denić et al. 2021)
- Correlational or causal?

From humans to language models

Why?

- If we're interested in learnability
- If we want have very detailed access to representations

We found something in a language model. So what?

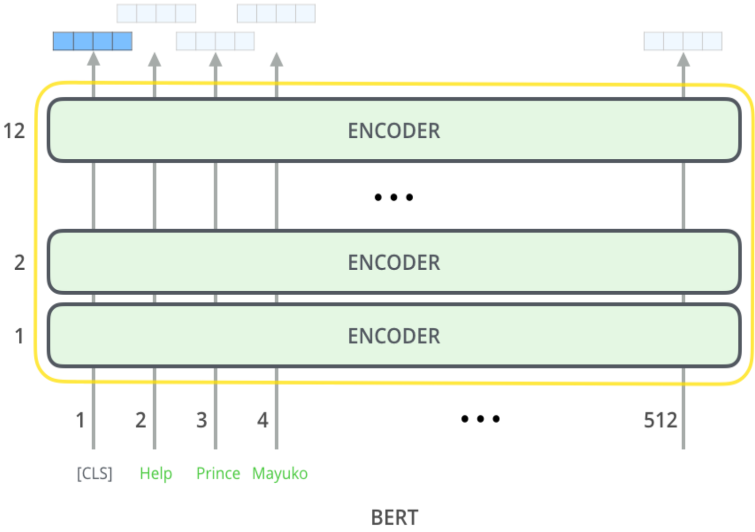
- This can be learned from text only
- Language models as 'algorithmic linguistic theories'

(Baroni 2021)

Large pre-trained language models

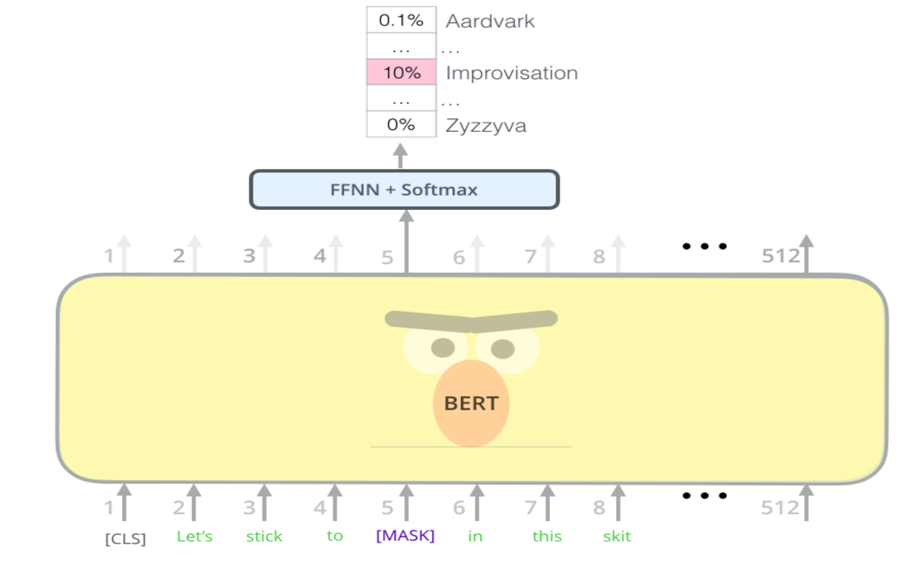
- Transformer architecture (Vaswani 2017) gave rise to a whole generation of SOTA NLP models, mainly:
  - BERT family of models (Devlin et al. 2019)
  - GPT family of models (Radford et al. 2019)
- Very large amount of training data
- **A lot** of trainable parameters
- Pretty good representations that are, roughly, task-agnostic (very adjustable to different tasks)

BERT: 1-minute intro



Picture from `https://jalammar.github.io/illustrated-bert/`

BERT: 1-minute intro



Picture from `https://jalammar.github.io/illustrated-bert/`

Part 1:
Polarity-sensitivity in monolingual BERT

(Bylinina and Tikhonov 2021a)

Polarity-sensitivity in pre-trained LMs: related work

- NPIs as part of combined benchmarks
  (Marvin & Linzen 2018; Hu et al. 2020)
- Main object of study
  (Jumelet & Hupkes 2018; Warstadt et al. 2019; Jumelet et al. 2021)
- Different set-ups: zero-shot, with fine-tuning, full training
- All these studies are monolingual (English)
- General conclusion – it's complicated (but not bad):
  neural models' recognition of polarity-sensitivity varies for different licensers and scope configurations

Logical vs. subjective polarity

Polarity via logical monotonicity	
NEG > AFF;	AT MOST > AT LEAST
NO > SOME;	AT MOST > BETWEEN / EXACTLY
FEW > MANY;	FEW > BETWEEN / EXACTLY
FEWER > MORE;	FEWER > BETWEEN / EXACTLY

Logical vs. subjective polarity

Polarity via logical monotonicity	
NEG > AFF;	AT MOST > AT LEAST
NO > SOME;	AT MOST > BETWEEN / EXACTLY
FEW > MANY;	FEW > BETWEEN / EXACTLY
FEWER > MORE;	FEWER > BETWEEN / EXACTLY

Subjective polarity / monotonicity	
NEG > AT MOST;	NO > FEW
NEG > FEW;	NO > FEWER
NEG > FEWER;	FEWER > AT MOST
NO > AT MOST;	EXACTLY > BETWEEN

Basic BERT experiment: Data and set-up (with omissions)

Synthetic data:

- Basic transitive template-generated sentences; filtered by GPT-2 perplexity; modified for different conditions
- 12 datasets 20k sentences each:
  AFF; NEG; SOME; NO; MANY; FEW; MORE THAN 5; FEWER THAN 5; AT LEAST 5; AT MOST 5; EXACTLY 5; BETWEEN 5 AND 10
- 2 datasets 8230 sentences each:
  SOMEBODY / SOMEONE / SOMETHING
  NOBODY / NO ONE / NOTHING

A girl crossed any roads.
A girl **didn't** cross any roads.
Some girls crossed any roads.
Somebody crossed any roads.

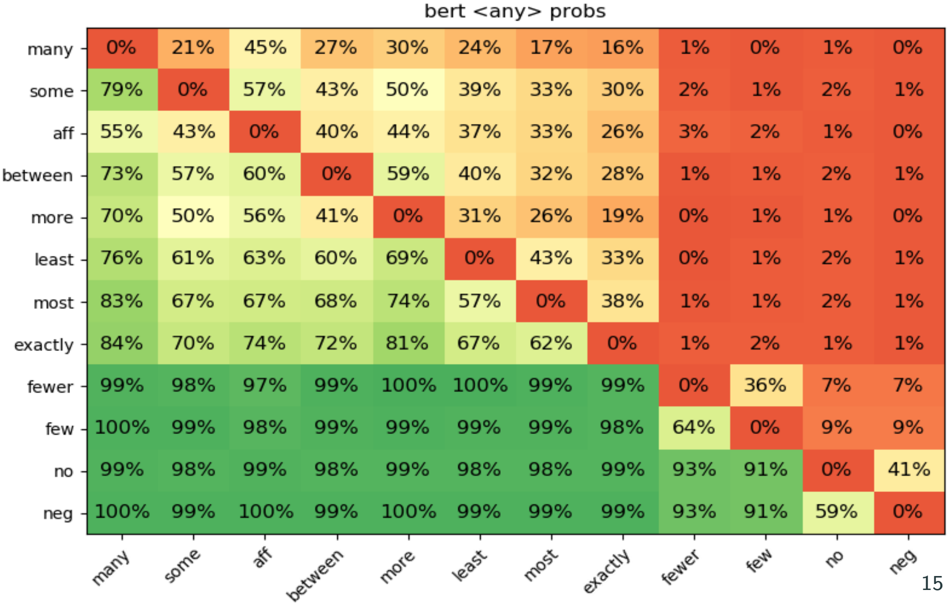
Basic BERT experiment: Data and set-up (with omissions)

$$\frac{\sum_{s \in D}[p(\texttt{[MASK]}=m|s_{cond\_i})>p(\texttt{[MASK]}=m|s_{cond\_j})]}{|D|}$$

$\langle$AFF, NEG$\rangle$: 5%

In 5% of the minimal pairs, the probability of an NPI in the affirmative sentence was higher than in its negative counterpart

BERT NPI results per licenser type



BERT NPI results

Polarity via logical monotonicity	
NEG > AFF; ✓	AT MOST > AT LEAST ✓
NO > SOME; ✓	AT MOST > BETWEEN / EXACTLY ✓
FEW > MANY; ✓	FEW > BETWEEN / EXACTLY ✓
FEWER > MORE; ✓	FEWER > BETWEEN / EXACTLY ✓

BERT NPI results

Polarity via logical monotonicity	
NEG > AFF; ✓	AT MOST > AT LEAST ✓
NO > SOME; ✓	AT MOST > BETWEEN / EXACTLY ✓
FEW > MANY; ✓	FEW > BETWEEN / EXACTLY ✓
FEWER > MORE; ✓	FEWER > BETWEEN / EXACTLY ✓

Subjective polarity / monotonicity	
NEG > AT MOST; ✓	NO > FEW ✓
NEG > FEW; ✓	NO > FEWER ✓
NEG > FEWER; ✓	FEWER > AT MOST ✓
NO > AT MOST; ✓	EXACTLY > BETWEEN ✓

Effect of cardinality: BERT experiment

Exactly two of the boxes contain anything.

??Exactly 98 of the boxes contain anything.

(Crnič 2014; Alexandropoulou et al. 2020)

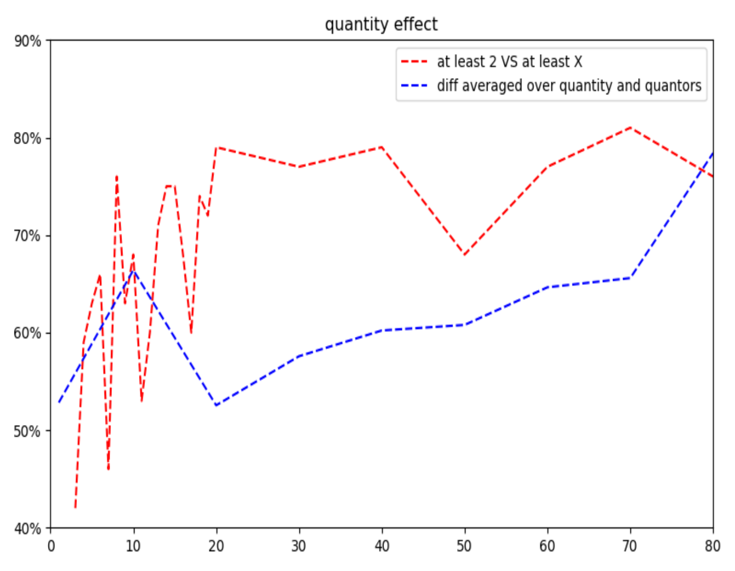
Effect of cardinality: BERT experiment

Exactly two of the boxes contain anything.
??Exactly 98 of the boxes contain anything.

(Crnič 2014; Alexandropoulou et al. 2020)

- Numerals $[2-20, 30, 40, 50, 60, 70, 80, 90]$
- As before, minimal pairs differing only in the numeral
  – testing all quantifiers containing numerals (*at least*, *at most*, *fewer than*, *more than*)

Effect of cardinality: BERT experiment



Effect of cardinality: humans

- **forced-choice task**, **2x2**:

  NUM: *five* vs. *seventy*; QUANT: *at least* vs. *more than*

- **6 test conditions**:

  at least five vs. *at least seventy*

  at least five vs. *more than five*

  at least five vs. *more than seventy*

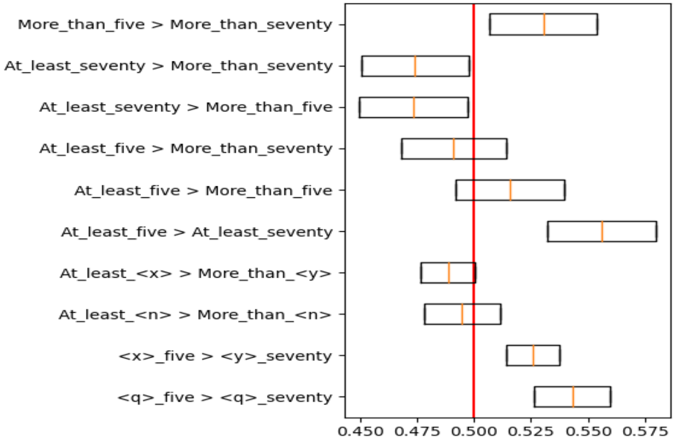
  at least seventy vs. *more than five*

  at least seventy vs. *more than seventy*

  more than five vs. *more than seventy*

- 50 patterns (out of 20k) give 2500 pattern pairs * 6 conditions = **15k unique test items**
- Each of the self-reported English-speaking participants recruited via Yandex.Toloka saw **38 pairs of sentences**: 22 filler/control items and 16 test items
- 656 participants (= 10496 test items; > 2/3 of our pool)

Effect of cardinality: humans



- binomial test; • boxes = 95% confidence interval
- cardinality does play a role

Closer look: attention head (6, 2)

We calculated attention from *any* to the quantifier for every layer and every attention head, averaged across sentences and sorted.

`[CLS] it felt odd without any wards on it . [SEP]`

`[CLS] do you have any brothers or sisters ? [SEP]`

`[CLS] if there ' d been any babies present , he ' d have been un ##sto ##ppa ##ble . [SEP]`

`[CLS] we are unable to identify any others who knew of the scheme at the time it was being considered . [SEP]`

Closer look: attention head (6, 2)

We calculated attention from *any* to the quantifier for every layer and every attention head, averaged across sentences and sorted.

[CLS] it felt odd without any wards on it . [SEP]

[CLS] do you have any brothers or sisters ? [SEP]

[CLS] if there ’ d been any babies present , he ’ d have been un ##sto ##ppa ##ble . [SEP]

[CLS] we are unable to identify any others who knew of the scheme at the time it was being considered . [SEP]

[CLS] exactly two games told any stories . [SEP]

[CLS] exactly ninety games told any stories . [SEP]

Interim summary and Part 2 outlook

- Monolingual (English) BERT shows polarity-sensitivity patterns similar to those in humans
- Generalizations to licensers beyond the basic set:
    - Cardinality effect (confirmed with humans)
    - Attention distribution impressionistically confirms this

Interim summary and Part 2 outlook

- Monolingual (English) BERT shows polarity-sensitivity patterns similar to those in humans
- Generalizations to licensers beyond the basic set:
  - Cardinality effect (confirmed with humans)
  - Attention distribution impressionistically confirms this

- What drives this generalization? Is it meaning-related?
- If yes, is it something that happens in natural language 'in general' (as in a 'statistical universal')?

$\star$ **Interventional tests + Multilingual models**

Part 2:
Polarity-sensitivity in multilingual language models

(Bylinina and Tikhonov 2021b)

Multilingual pre-trained models

Multilingual BERT (mBERT) (Devlin et al. 2019)

- 104 languages
- Token vocabulary: 110k shared tokens
- Training data: Entire Wikipedia dumps for the 104 languages

Both models:

- Main training objective: masked token prediction
- Lower-resource languages are upweighted in sampling
- No input language marker or language encodings
  (to facilitate code-switching and adding new languages)

Multilingual pre-trained models

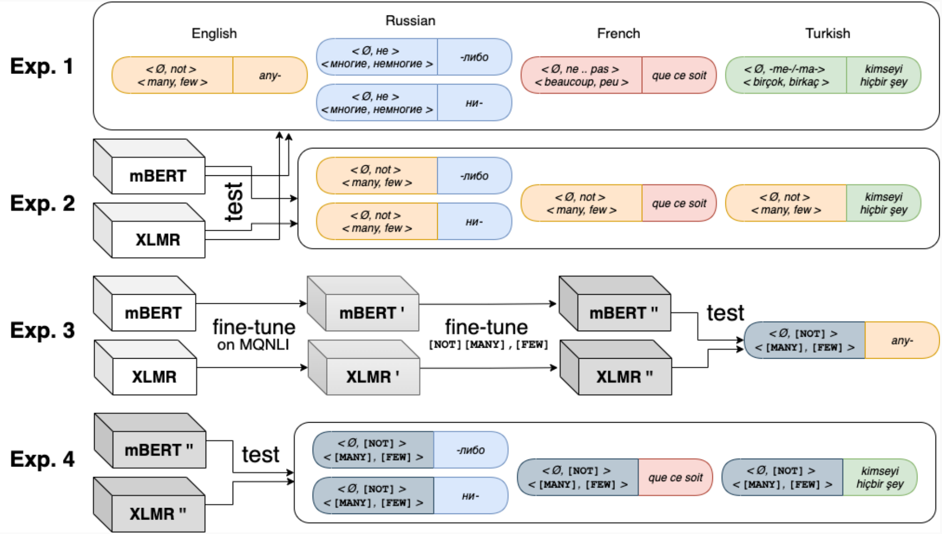
XLM-RoBERTa (XLMR) (Conneau et al. 2019)

- 100 languages
- Token vocabulary: 250k shared tokens
- Training data: CommonCrawl corpus

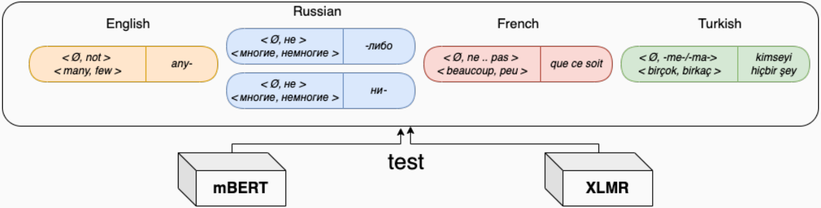
Both models:

- Main training objective: masked token prediction
- Lower-resource languages are upweighted in sampling
- No input language marker or language encodings
  (to facilitate code-switching and adding new languages)

Experiments (Bylinina & Tikhonov 2021b)



Experiment 1: Data and procedure



	NPI	NEG	MANY	FEW
EN	anything / anybody	not	many	few
FR	quoi que ce soit / qui que ce soit	ne ... pas	beaucoup	peu
RU	ничто / никто что-либо / кто-либо	не	многие	немногие
TR	hiçbir şey / kimseyi	-me- / -ma-	birçok	birkaç

Experiment 1: Data and procedure cont'd

Synthetic datasets generated with a pattern and filtered by GPT-2 perplexity. 10k quadruples per language:

$$\langle \text{ AFF, NEG, MANY, FEW} \rangle$$

The letters meant anything.
The letters **did not** mean anything.
Many letters meant anything.
Few letters meant anything.

Pair-wise comparison $\langle\text{AFF, NEG}\rangle$, $\langle\text{ MANY, FEW}\rangle$

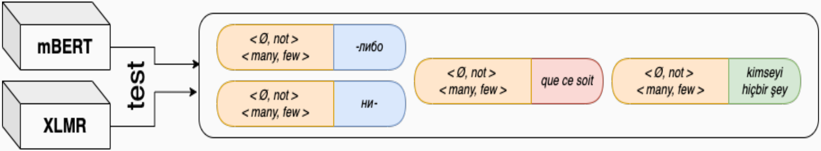
Same proportional metric as before:

$$\frac{\sum_{s \in D}[p(\texttt{[MASK]}=m|s_{cond\_i})>p(\texttt{[MASK]}=m|s_{cond\_j})]}{|D|}$$

Experiment 1: Results

	⟨AFF,NEG⟩		⟨MANY,FEW⟩	
	mBERT	XLMR	mBERT	XLMR
en	0.45%	0.35%	20.45%	25.27%
fr	4%	37.1%	20.42%	32.93%
ru ни-	0.12%	0.17%	20.66%	21.46%
ru -либо	21.92%	35.96%	46.74%	12%
tr	18.12%	13.99%	45.23%	30.11%

Experiment 2: Data and procedure



EN licensers (*not*, *many*, *few*) transplanted into sentences from other languages →
Polarity interactions across a language boundary.

Few люди ничего потеряли.
few people anything lost

Procedure: exactly the same as in Exp. 1.

Experiment 2: Results

	⟨AFF,NEG⟩		⟨MANY,FEW⟩	
en+	mBERT	XLMR	mBERT	XLMR
fr	1.28%	4.71%	44.73%	21.21%
ru ни-	23.09%	13.63%	37.36%	49.87%
ru -либо	45.6%	0.35%	13.41%	25.27%
tr	44.43%	33.75%	52.94%	62.48%

Experiment 3: Data and procedure

Artificial language learning

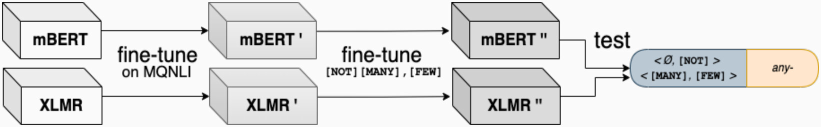
(Friederici et al. 2002; Finley & Badecker 2009; Culbertson et al. 2012; Ettlinger et al. 2014; Kanwal et al. 2017; Motamedi et al. 2019)

- a fragment of an artificial language: expressions that do not belong to the participants' language;
- **training phase**: information about the language fragment is given to participants (property A);
- **test phase**: checking what other knowledge, beside the provided, was inferred during training (property B)

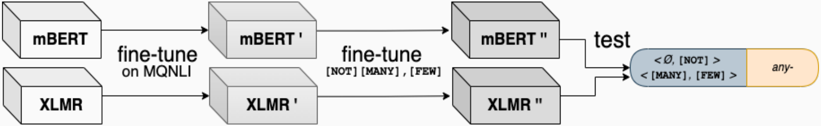
In the context of pre-trained LMs:

Thrush et al. 2020; Bylinina, Tikhonov & Garmash 2021.

Experiment 3: Data and procedure



Experiment 3: Data and procedure



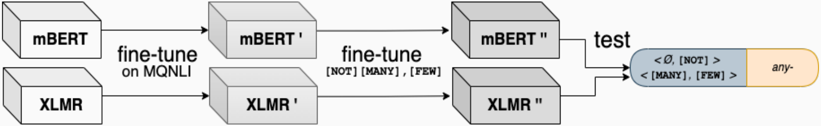
MQNLI (Geiger et al. 2020, 2021):

- Template-generated sentences
- Entailment labels assigned using 'natural logic' rules

Q_s	Adj_s	N_s	Neg	Adv	V	Q_o	Adj_o	N_o
every	angry	philosopher	doesn't		draw	some		doors
every		philosopher		honestly	draws	some	Irish	doors

`contradiction`

Experiment 3: Data and procedure



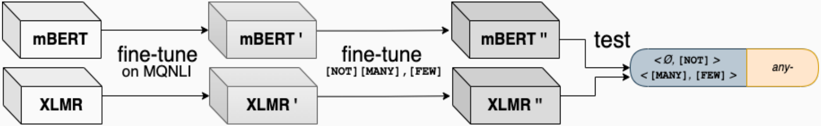
MQNLI (Geiger et al. 2020, 2021):

- Template-generated sentences (500k pairs)
- Entailment labels assigned using 'natural logic' rules

Q_s	Adj_s	N_s	Neg	Adv	V	Q_o	Adj_o	N_o
every		milkman	[NOT]	stylishly	pats	not every		helmet
every	jealous	milkman	[NOT]		pats	some	flexible	helmet

entailment

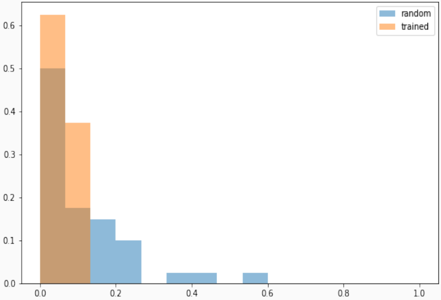
Experiment 3: Data and procedure



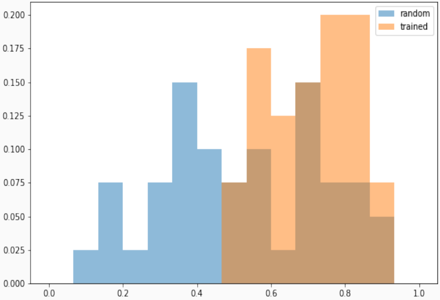
- **Stage 1:** NLI fine-tuning on a fragment of original MQNLI (20k training items + 3.5k val+test, no lexical overlap)
- **Stage 2:** NLI fine-tuning of the Stage 1 output model with modified MQNLI items ([NOT] as negation; [FEW] as a DE quantifier; [MANY] as a UE quantifier) (16k items total, 80:10:10 train:val:test). Repeat 40 times, reshuffling data and with different random initialisations: **40 new triples**
- Transplant trained tokens into original models for evaluation
- Evaluation as before + comparison to a random baseline

Experiment 3: mBERT results

$\langle$AFF, [NOT]$\rangle$ vs. $\langle$AFF, RAND$\rangle$

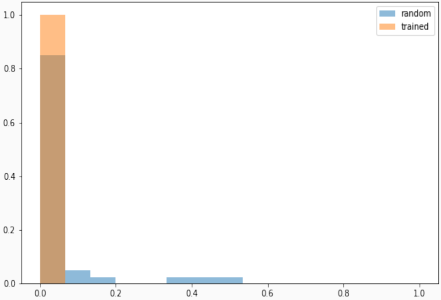


$\langle$[MANY], [FEW]$\rangle$ vs. $\langle$RAND, RAND$\rangle$

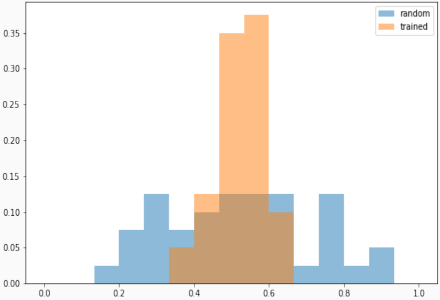


Experiment 3: XLMR results

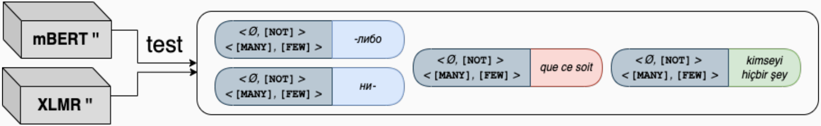
$\langle$AFF, [NOT]$\rangle$ vs. $\langle$AFF, RAND$\rangle$



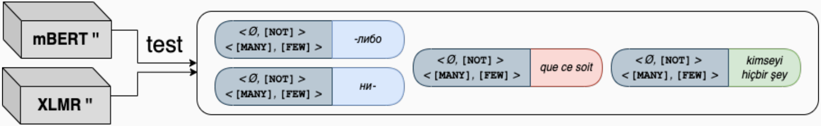
$\langle$[MANY], [FEW]$\rangle$ vs. $\langle$RAND, RAND$\rangle$



Experiment 4: Data and procedure



Experiment 4: Data and procedure



- Like in Exp. 2, we make hybrid sentences, transplanting new tokens into French, Russian and Turkish items:

  `[FEW]` d'amis ont vu quoi que ce soit

  few of.friends have seen anything

- Measure polarity interaction in the same way as in Exp. 2, but against a random baseline, like in Exp. 3

Experiments 3 & 4: mBERT results

contrast	lang	rand_mean	tr_mean	mw pval
aff>neg	En	0,124	0,056	0,61%
	Ru *ни*	0,798	0,858	19,38%
	Ru *либо*	0,86	0,823	0,073%
	Fr	0,262	0,315	9,98%
	Tr	0,831	0,8294	33,34%
many>few	En	0,518	0,708	0,01%
	Ru *ни*	0,541	0,61	77,65%
	Ru *либо*	0,519	0,597	21,27%
	Fr	0,526	0,639	10,39%
	Tr	0,552	0,488	29,19%

Experiment 4: XLMR results

contrast	lang	rand_mean	tr_mean	mw pval
aff>neg	En	0,053	0,006	0,02%
	Ru *ни*	0,357	0,058	0,00000008%
	Ru *либо*	0,776	0,684	0,00000026%
	Fr	0,594	0,408	0,000000001%
	Tr	0,744	0,562	0,000000000%
many>few	En	0,519	0,529	55,07%
	Ru *ни*	0,511	0,563	7,5%
	Ru *либо*	0,515	0,566	13,2%
	Fr	0,496	0,536	3,63%
	Tr	0,48	0,563	0,02%

Conclusions and future work

- mBERT and XLMR do a decent job encoding polarity-sensitivity in languages we checked
- The polarity-based interaction mechanism is partly cross-linguistically general (speculation: depending on how structurally similar languages are)
- Polarity-sensitivity is meaning-driven: found for negation but not for quantifiers

Conclusions and future work

- mBERT and XLMR do a decent job encoding polarity-sensitivity in languages we checked
- The polarity-based interaction mechanism is partly cross-linguistically general (speculation: depending on how structurally similar languages are)
- Polarity-sensitivity is meaning-driven: found for negation but not for quantifiers

- What happened with quantifiers?
- NPI-licensing by random tokens in English but not in other languages – what's up with that?

Conclusions and future work

- mBERT and XLMR do a decent job encoding polarity-sensitivity in languages we checked
- The polarity-based interaction mechanism is partly cross-linguistically general (speculation: depending on how structurally similar languages are)
- Polarity-sensitivity is meaning-driven: found for negation but not for quantifiers

- What happened with quantifiers?
- NPI-licensing by random tokens in English but not in other languages – what's up with that?

Thank you!