Mareike Hartmann[1]
Miryam de Lhoneux[2,3,4]
Daniel Hershcovich[2]
Yova Kementchedjhieva[2]
Lukas Nielsen[2]
Chen Qiu[5]
Anders Søgaard[2]

A Multilingual Benchmark for Probing Negation-Awareness with Minimal Pairs

[1]German Research Center for Artifical Intelligence (DFKI), Germany. [2]University of Copenhagen, Denmark.
[3]Uppsala University, Sweden.[4]KU Leuven , Belgium. [5]Wuhan University of Science and Technology, China

We generate NLI probing datasets based on minimal pairs in 5 languages.

Minimal Pairs

Premise	Hypothesis	
He was **not** a nice man.	He was the nicest man you'll ever meet!	✗ C
He was ~~not~~ a nice man.	He was the nicest man you'll ever meet!	? N
She was **not** impressed by the signs.	It was certain that she saw the signs.	✓ E
She was ~~not~~ impressed by the signs.	It was certain that she saw the signs.	✓ E

Important negation
Is the model aware of negation?

Unimportant negation
Does the model exploit negation as lexical cue?

Creating the Datasets

1 Compile lists of negation cues

🇬🇧 no, not, never, nobody, without,...
🇧🇬 не, никога не, няма. никой не, нямаше, ...
🇩🇪 nicht, keine, nie, nichts, niemand, ...
🇫🇷 ne pas, jamais, aucun, rien, ne plus, ...
🇨🇳 不, 没, 未, 没有, 从来没有, ...

2 Match XNLI examples

Match examples with
at most one cue **in premise** and **at most** one cue in **hypothesis**

✓ P: I do **not** own a bike. H: I do **not** own a car.
✗ P: **Never** mind, I do **not** own a bike. H: I own a bike.

3 Rewrite or discard

✓They have written ~~anything~~ something about it.
✗ ~~Never~~ mind the question about it.
✗ My friend is deaf, so he can~~not~~ listen to music.

4 Relabel

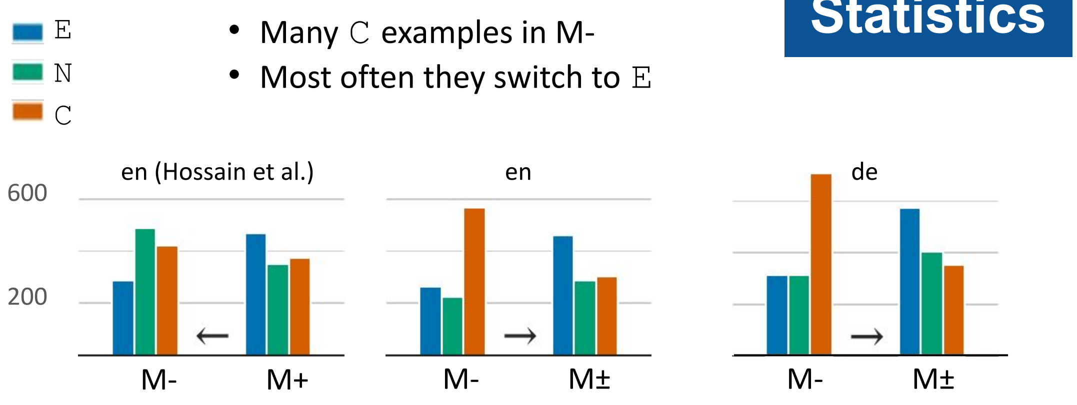
P: They were ~~not~~ impressed by the signs.
H: It was certain that they saw the signs.
✓ E → ✓ E

Adding vs Removing Negation

- We build our minimal pairs by **removing** negation
  M- ⋈ M±
- Hossain et al. (2020) **add** negation
  by inserting not to negate the main verb
  M+ ⋈ M- (English only)

Statistics

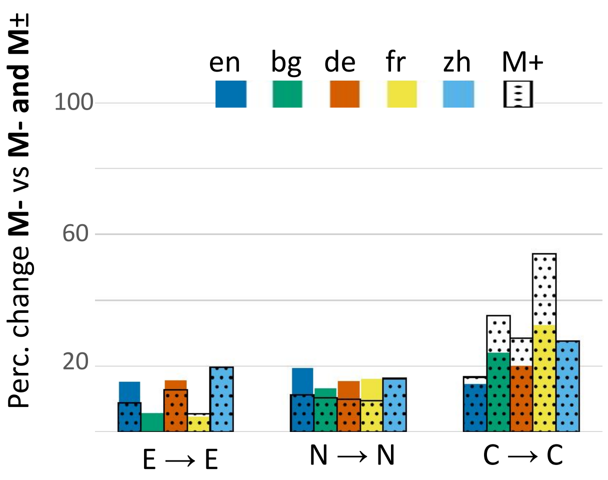
- Many C examples in M-
- Most often they switch to E



We probe mBERT fine-tuned for NLI on our datasets.

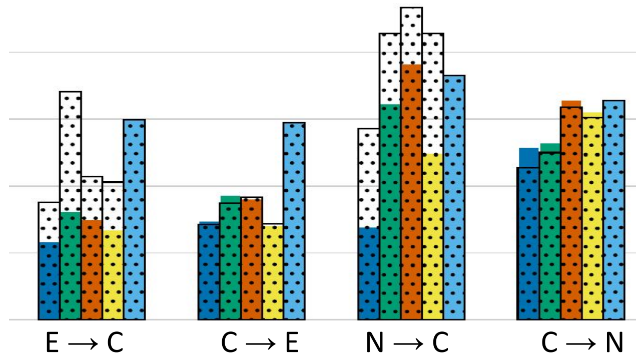
Probing mBERT

Unimportant Negations



- Negation (mismatch) indicates C class
  Dasgupta et al. (2018)
  Poliak et al. (2018)
  Gururangan et al. (2018)
  McCoy and Linzen (2019)
- Bias transfers across languages

Important Negations



- E↔C switch is easier than N↔C
- Possible explanation: high Pr.-Hypo. overlap for E and C classes
  Dasgupta et al. (2018)
  Naik et al. (2018)
  McCoy et al. (2019)