A Multilingual Benchmark for Probing Negation-Awareness with Minimal Pairs

Mareike Hartmann[1*], **Miryam de Lhoneux**[2,3,4], **Daniel Hershcovich**[2],
Yova Kementchedjhieva[2], **Lukas Nielsen**[2], **Chen Qiu**[5], **Anders Søgaard**[2]
[1]German Research Center for Artificial Intelligence (DFKI), Germany
[2]University of Copenhagen, Denmark [3]Uppsala University, Sweden [4]KU Leuven, Belgium
[5]Wuhan University of Science and Technology, China
mareike.hartmann@dfki.de, chen@wust.edu.cn
{ml,dh,yova,lukas.christian.nielsen,soegaard}@di.ku.dk

Abstract

Negation is one of the most fundamental concepts in human cognition and language, and several natural language inference (NLI) probes have been designed to investigate pretrained language models' ability to detect and reason with negation. However, the existing probing datasets are limited to English only, and do not enable controlled probing of performance in the absence or presence of negation. In response, we present a multilingual (English, Bulgarian, German, French and Chinese) benchmark collection of NLI examples that are grammatical and correctly labeled, as a result of manual inspection and editing. We use the benchmark to probe the negation-awareness of multilingual language models and find that models that correctly predict examples with negation cues often fail to correctly predict their counter-examples *without* negation cues, even when the cues are irrelevant for semantic inference.

1 Introduction

Negation is a fundamental concept of human cognition, for asserting the falsity of a proposition (Heinemann, 2015). The linguistic markers of negation enable us, for example, to deny that events happened, or to express the absence of objects (Horn, 1989). Hence, correctly processing markers of negation is a key building block of language comprehension, often acquired by children in multiple stages (Thornton and Tesan, 2013).

There is ample evidence that natural language processing (NLP) models struggle with processing negation: Negation has been identified as a frequent source of error for various NLP tasks, such as sentiment analysis (Barnes et al., 2019), statistical and neural machine translation (Fancellu and Webber, 2015; Hossain et al., 2020a; Bentivogli et al., 2016), and question answering (Staliūnaitė and Iacobacci, 2020).

To complement insights from anecdotal error analysis, several diagnostic datasets have recently been designed to explicitly investigate the negation processing capabilities of pretrained language models, either directly on the encoder (Ettinger, 2020; Kassner and Schütze, 2020) or after fine-tuning the model for downstream tasks such as natural language inference (NLI) (Naik et al., 2018; Kim et al., 2019; Geiger et al., 2020; Richardson et al., 2020; Hossain et al., 2020b) or sentiment analysis (Li and Huang, 2009; Zhu et al., 2014). These datasets have been useful to shed light on models' negation processing capabilities, but the diagnostic datasets so far only cover English. With our work, we extend the efforts of analyzing a model's negation awareness to a multilingual setup, by deriving an analysis dataset from the multilingual XNLI dataset (Conneau et al., 2018).

The goal of our work is to provide multilingual datasets to determine if models adequately process negation. We not only want to know if they make correct inferences in the presence of negation but also if correct inferences are a result of adequately modeling the semantics of the negation, or if they are a result of merely exploiting shallow heuristics. In our datasets, we measure this negation awareness by quantifying the extent to which: (i) models correctly change their prediction when a label-changing negation is inserted/removed from an input sentence; and (ii) models correctly keep their predictions unchanged when a label-preserving negation is inserted/removed from an input sentence. To this end, we create *minimal pairs* of NLI examples, that only differ in the presence/absence of negation (see Table 1 for examples), and hence allow a detailed investigation of a model's reaction to the phenomenon of negation. We evaluate a multilingual language model on our new benchmark, and find that even though it correctly predicts exam-

[*] The research was carried out while the author was employed at the University of Copenhagen.

Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL), pages 244–257
November 10–11, 2021. ©2021 Association for Computational Linguistics

(1) C	P: My grandfather was **not** a nice man H: My grandpa was the nicest guy you'll ever meet!	→	N	P: My grandfather was a nice man H: My grandpa was the nicest guy you'll ever meet!
(2) E	P: The rabbis were **not** impressed by these signs H: It was certain that the rabbis saw the signs.	→	E	P: The rabbis were impressed by these signs H: It was certain that the rabbis saw the signs.

Table 1: Label-changing (1) and label-preserving (2) negation removal of the negation cue marked in bold. C, N, and E, refer to the NLI relation categories contradiction, neutral, or entailment, that hold between the premise P and the hypothesis H.

ples with negation cues, it often fails to correctly predict their counter-examples *without* negation cues, even when the cues are irrelevant for semantic inference.

Contributions We create five new targeted probing datasets in English, French, German, Bulgarian and Chinese.[1] Our probing datasets consist of minimal pairs of NLI examples, with manual annotations that allow us to measure the extent of a model's negation awareness. In contrast to other negation probing datasets, our minimal pairs allow for a detailed comparison of a model's performance on important (label-changing) and unimportant (label-preserving) negations. We find that in a large amount of cases, the model cannot correctly adjust its prediction when the negation is removed, pointing towards the model's inability to model the negation's effect on sentence semantics.

2 Related Work

The processing of negation has been investigated in several works, either by creating dedicated diagnostic datasets, or by investigating the phenomenon as part of a more general analysis. An overview over existing datasets can be found in Table 2.

Crafting adversarial NLI examples based on unimportant negations Several works have established that negation cues in NLI examples are a strong indicator for a contradiction label (Gururangan et al., 2018; Dasgupta et al., 2018; Poliak et al., 2018). Hence, negation can be used to craft adversarial NLI examples, i.e., to apply a label-preserving change that makes the model incorrectly change its prediction, by either inserting or removing unimportant negations. Such adversarial examples have been derived from the MultiNLI (MNLI) dataset (Williams et al., 2018), a large English dataset with NLI examples from multiple genres. Naik et al. (2018) automatically build such adversarial examples by adding a tautology (*and false is not true*) to the end of every hypothesis in the MNLI data, leading to a drop in model accuracies caused by a large amount of false positives (FPs) for the neutral class, rather than an expected increase in FPs for the contradiction class. The authors hypothesize that this pattern stems from decreasing lexical overlap by adding the tautology, and as low lexical overlap is indicative for the neutral class, the model mispredicts the example as neutral. Aspillaga et al. (2020) confirm that also transformer-based models perform poorly on this negation stress test. On their challenge dataset for non-entailed subsequences based on the MNLI data, McCoy and Linzen (2019) find that models exploit a mismatch between negation cues in premise and hypothesis to predict non-entailment, and that removing unimportant negation cues decreases model accuracy to almost 0.

Measuring negation awareness based on important negations Instead of focusing on unimportant negations in order to uncover lexical biases in the data, other works focus on processing important negations, i.e., negations that induce a label change when added/removed. Kim et al. (2019) build an NLI dataset based on MNLI premise-hypothesis pairs that contain antonyms. They build combinations of syntactically negated and non-negated versions of the premise and hypothesis, and re-label the resulting examples. Their results confirm that models perform worse on this analytic dataset than on the original dataset, and that models are worse in predicting pairs with syntactic negation than pairs with antonyms. Hossain et al. (2020b) create a dataset containing challenging negations by adding the syntactic negation cue *not* to the main verb of the premise and/or the hypothesis of MNLI training examples. They find that their newly created examples are more challenging for the model than the original negated examples.

[1] The datasets are available at https://github.com/mahartmann/negationminpairs.

Authors	Task	Base Dataset	Data Creation	Langs
Ettinger (2020)	MLM (cloze)	Psycholinguistic stimuli	-	en
Kassner and Schütze (2020)	MLM (cloze)	-	Template filling	en
Naik et al. (2018)	NLI	MNLI	Adding tautology	en
Kim et al. (2019)	NLI	MNLI	Inserting/removing negation, swapping antonyms	en
Hossain et al. (2020b)	NLI	MNLI	Inserting negation	en
Richardson et al. (2020)	NLI	-	Template filling	en
Geiger et al. (2020)	NLI	SNLI	Replacing single words	en
Ours	NLI	XNLI	Removing negation to build minimal pairs	en, bg, de, fr, zh

Table 2: Our approach (bottom line) in comparison to related work on diagnostic datasets involving negation.

Other probing datasets for negation The probes described above study negation as present in existing NLI examples. The following works are interested in specific inference mechanisms, and artificially create data requiring the inference to be recognized correctly. Geiger et al. (2020) investigate if models can learn interactions between lexical entailment and negation, in particular the algorithm behind downward monotonicity (e.g., *dance* entails *move*, and *not move* entails *not dance*), and find that models cannot solve the task when fine-tuned on MNLI, but when fine-tuned on the challenge dataset. Their dataset is created by substituting single words in examples from the SNLI dataset (Bowman et al., 2015). In contrast, Richardson et al. (2020) use a template to build a probing dataset with syntactic negation of verbs, that requires lexical inference and reasoning skills. Again, models fine-tuned on the standard MNLI data perform poorly, but improve when fine-tuned on the target dataset. Instead of fine-tuning on an NLI task, Ettinger (2020) and Kassner and Schütze (2020) directly probe pre-trained encoders using a cloze language modeling task. They find that a language model makes the same predictions in negative and assertive contexts, but it is unclear to what extent we can expect a language model to learn the semantics of negation from an unsupervised pretraining task (see Bender and Koller (2020)).

Our multilingual benchmark differs from previous work on probing negation-awareness, not only in being multilingual, but also in using minimal pairs. Minimal pairs have been used for other diagnostic datasets, e.g. to check if language models assign higher probability to grammatical sentences (Marvin and Linzen, 2018; Warstadt et al., 2020). Gardner et al. (2020) suggest to augment existing datasets with minimally different examples to test a model's decision boundaries.

With our multilingual benchmark, we contribute to a line of research that probes a model's handling of linguistic phenomena in multiple languages, including features of sentence representations (Ravishankar et al., 2019b,a), tenses (Li and Wisniewski, 2021), gender bias (González et al., 2020), numerical understanding (Johnson et al., 2020), and lexical semantics (Vulić et al., 2020).

3 Approach

Our goal is to verify if a model adequately processes negation: In the presence of negation, the model should make a correct prediction. The most naive way to look at this is to check model performance on examples that contain negations. However, even if the model correctly predicts such examples, it might do so for the wrong reasons: It might (1) ignore the negation completely or not properly model its effect on the inference category. In such cases, it would have predicted the same class if the negation was absent. In order to check for this deficiency, we need minimal pairs with relevant or *important negations*, i.e. the presence/absence of negation changes the correct inference label. On the other hand, it might (2) predict an inference label based on the presence/absence of negation in the sentence, no matter if the negation is relevant to the inference relation or not. This could either be because of the established biases related to negation cues, or because the model does not correctly model the scope of the negation (assign a wrong scope, or does not have a notion of scope at all). In order to check for this deficiency, we need minimal pairs with irrelevant or *unimportant negations*, i.e. the presence/absence of the negation does not change the correct inference label. Considering minimal pairs with both types of negation allows us to gain a more complete insight into model behaviour and sheds light on the questions if the model (1) is aware of negation and its effect on sentence semantics and (2) models the effect of negation beyond its mere presence/absence in a sentence. Overall, our datasets contribute to

✓	Newsweek has **never** written anything about the Hamptons.	→	Newsweek has written *something* about the Hamptons.
✓	Sie werden für **nichts** im Voraus zahlen. *They will pay for nothing in advance*	→	Sie werden für *etwas* im Voraus zahlen. *They will pay for something in advance*
✗(1)	**Never** mind the question of whether the Dow Jones industrial average is the proper measure.		
✗(2)	Mein Freund ist taub, also kann er **keine** Musik hören. *My friend is deaf, so he cannot listen to music*	→	Mein Freund ist taub, also kann er Musik hören. *My friend is deaf, so he can listen to music*
✗(Mismatch)	P: Oh, ist es das, worüber du redest. *Oh, that is what you talk about.* H: Du rufst nicht von dort an. *You are not calling from there*		Original P: oh is that where you're talking from Original H: You are not calling from there

Table 3: Examples of accepted rewritten (✓) and discarded (✗) sentences after removing the negation cue. Reasons for discarding are (1) no easy rewrite possible, (2) negation removal leads to a sentence that is contradictory in itself. The bottom premise-hypothesis example is discarded, as the gold label `contradiction` of the original EN example does not match with its DE translation.

understanding if a model adequately represents a negation's effect on sentence semantics.

Desired properties of our benchmark We want to create a multilingual benchmark for studying negation-awareness derived from existing NLI examples. In contrast to previous work, we want to focus on a more varied set of negation cues, and consider important and unimportant negations at the same time. Instead of individual examples, our analysis will be based on minimal pairs that only differ in the presence/absence of negation, and allow us to directly investigate the effect of a negation cue on the model prediction. In order to create a benchmark with these properties, we proceed in the following steps: we extract sentences that contain negation cues in the XNLI data based on lists of negation cues, and build minimal pairs by removing the negation cue. We then manually adjust the grammar of the sentence if necessary, and relabel the new example. Based on the labels of the original and modified examples, the minimal pair can be classified as differing in an important or unimportant negation cue.

Multilingual negation cues We extract a list of English (EN) negation cues from datasets annotated for the negation scope resolution task, in particular from the Sherlock dataset (Morante and Daelemans, 2012), as a starting point to automatically compile such lists for French (FR), German (DE), and Bulgarian (BG). We apply the `fast_align` alignment tool (Dyer et al., 2013) to word-align English and target language sentences sourced from EuroParl (Tiedemann, 2012). Based on a many-to-many alignment, we extract the twenty most common translations in the target language. The resulting lists were refined by native speakers of each language, by removing items that would not be considered negation cues, and completing negation cues that were only partially translated. The final lists comprise syntactic, lexical and morphological negation cues. For Chinese (ZH), we directly extract cues from the Chinese version of the Sherlock dataset (Liu et al., 2018).

Deriving minimal pairs Given the translated negation cues, we identify negated sentences in the EN, ZH, DE, FR, and BG XNLI datasets (Conneau et al., 2018), and use them as a base to build minimal pairs.[2] The XNLI dataset comprises development and test splits in 15 languages. The English part was collected following the same setup as for the MNLI dataset (Williams et al., 2018), and the non-English parts are manual translations of the English data. As we want to study the effect of negation removal, we focus on sentences that contain exactly one negation marker in either premise and/or hypothesis. We remove the matched negation cues and manually verify the sentence (see Table 3 for examples), i.e., we re-write the sentence to adjust for negation removal grammatically, and discard it if (i) adjusting the sentence requires a complex re-write; (ii) removing the negation leads to a contradictory statement; or (iii) translation quality is low. We then pair the verified sentences back into their original premise-hypothesis pairs, and re-label the resulting new NLI example.[3] In

[2] We derive our minimal pairs from the test splits, removing examples for which the assigned gold label does not correspond to the majority label assigned by the five annotators.

[3] All annotations were done by native speakers of the respective languages and with expert knowledge of linguistics; except for EN, for which the annotator is a C2 level speaker.

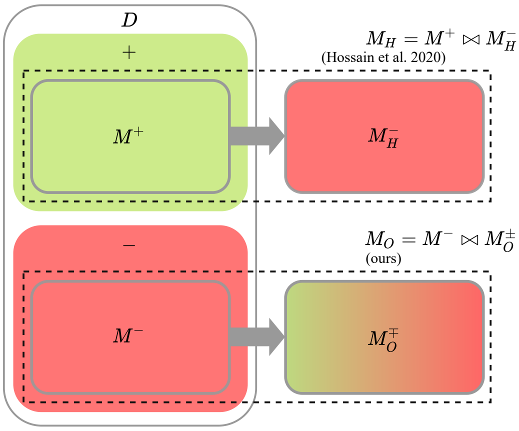


Figure 1: Deriving minimal pairs by removing negation cues (lower part) or adding negation cues (upper part) from original NLI examples with (M$^-$) or without (M$^+$) negation cues.

EN	not, no, never, nobody, without
BG	не *(no/not)*, никога не *(never)*, няма *(doesn't have/there isn't/won't)*, никой не *(nobody)*, нямаше *(*няма*, past tense)*
DE	nicht *(not)*, keine *(no)*, nie *(never)*, nichts *(nothing)*, niemand *(nobody)*
FR	ne pas *(not)*, jamais *(never)*, aucun *(no)*, rien *(nothing)*, ne plus *(no more)*
ZH	不*(not)*, 没*(without)*, 未*(not)*, 没有*(without)*, 从来没有*(never)*

Table 4: Most frequent negation cues per language.

	en	fr	de	bg	zh
original	0.36	0.54	0.39	0.45	0.42
modified	0.37	0.54	0.40	0.45	0.43

Table 5: Probability score for original and modified data measured with mBERT.

this step, we exclude numerous pairs for which the translation from EN had led to a mismatch with the original gold label.[4]

The datasets of minimal pairs resulting from our approach are visualized in the lower half of Figure 1. We start with a subset M$^-$ of the original XNLI data D, in particular the subset that contains exactly one of the target negation cues in either premise and/or hypothesis. By removing the negation cue, we derive M$_O^\pm$, the counterpart of the minimal pair after the target negation cue has been removed. Note that these examples can still contain a negation cue in either premise or hypothesis (if the original pair contained a negation cue in both).

Minimal pairs from Hossain et al. (2020b) In addition to our multilingual minimal pairs described above, we also derive a set of English minimal pairs from Hossain et al. (2020b)'s challenge dataset. Given that they derived new NLI examples by *inserting* negation (syntactic negation of the main verb using negation cue **not**), comparing the two sets of minimal pairs can help us to understand what different insights can be gained depending on whether negation is removed or inserted from the original examples. We map their modified examples back to their original counterparts, which come from the English training split of the MNLI dataset. This results in the set of minimal pairs visualized in the upper part of Figure 1, consisting of original NLI examples M$^+$ without any negation cues, paired with their modified counterparts in M$_H^+$ by inserting negation cues.

4 Data Analysis

Our new multilingual benchmark comprises NLI examples containing a range of different negation cues. The five most frequent cues are listed in Table 4, (a full list can be found in Appendix B). Across all languages, syntactic cues are the most frequent.

Naturalness of new examples The new premise-hypothesis pairs in M$_O^\pm$ only differ from their counterparts in M$^-$ in the absence of one negation cue, and minimal changes to adjust syntax after negation removal. This might lead to unnatural sentences, e.g. in the case of syntactic constructions that are typically only used in the presence of negation but uncommon without it. In order to rule this out as a possible factor for performance drop on the modified examples, we compare the probability mBERT (Devlin et al., 2019) assigns to utterances in the original data (M$^-$) and in the newly created data (M$_O^\pm$). We obtain a probability score for a sentence by masking one subword token at a time, extracting the prediction probability for the target subword token from the language model head, and averaging these prediction probabilities for all subword tokens (Wang et al., 2019). Results are shown in Table 5 and indicate that there is little change in the probability assigned to the original utterances and the newly created ones.

Label distribution Figure 2 shows how the class distribution changes between data with and without negation in both the minimal pairs derived

[4] For a detailed description of translation issues that we observed in the XNLI data, see Appendix A.

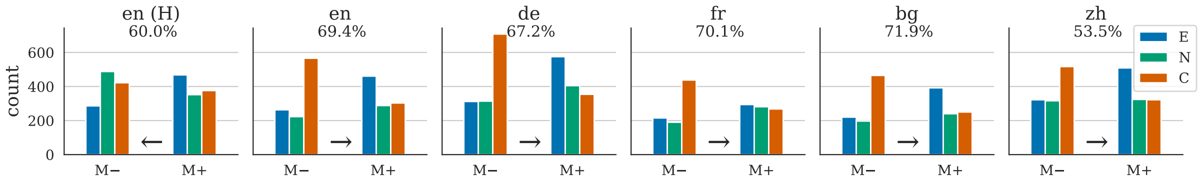
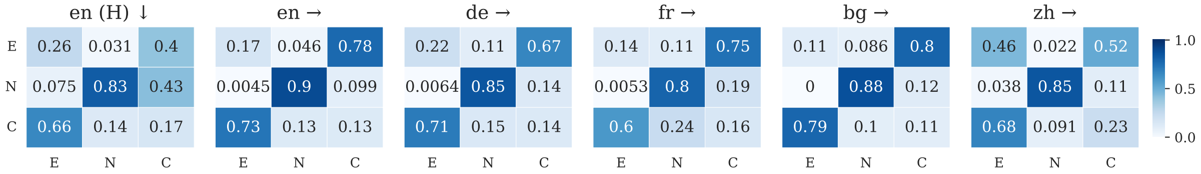


Figure 2: Number of pairs in each class with and without negation. The percentage indicates the rate of important negations. Arrows indicate the direction of modification.



en (H) ↓

	E	N	C
E	0.26	0.031	0.4
N	0.075	0.83	0.43
C	0.66	0.14	0.17

en →

	E	N	C
E	0.17	0.046	0.78
N	0.0045	0.9	0.099
C	0.73	0.13	0.13

de →

	E	N	C
E	0.22	0.11	0.67
N	0.0064	0.85	0.14
C	0.71	0.15	0.14

fr →

	E	N	C
E	0.14	0.11	0.75
N	0.0053	0.8	0.19
C	0.6	0.24	0.16

bg →

	E	N	C
E	0.11	0.086	0.8
N	0	0.88	0.12
C	0.79	0.1	0.11

zh →

	E	N	C
E	0.46	0.022	0.52
N	0.038	0.85	0.11
C	0.68	0.091	0.23

Figure 3: Label change between an example with (y-axis) and without (x-axis) negation. Arrows indicate the axis of normalization, which follows the direction in which minimal pairs were created.

using our approach (M_O) and the minimal pairs derived from Hossain et al. (2020b) (M_H). For M_O, we see the expected bias in M^- where the contradiction label occurs at a substantially higher rate than either of the other two classes. This bias is not observed in M_H, presumably because those negated examples were created heuristically. Across all subsets we see a shift in class distribution between original and modified instances indicative of important negations. Figure 3 sheds light on how the inference class changes based on presence/absence of the negation cue. The largest weight in most cases lies along the antidiagonal, indicating that negation often has the effect of reversing an entailment to a contradiction and vice versa, whereas the neutral class is fairly stable.Interestingly, adding negation to examples in M_H does not have quite the same clear effect, mostly due to the contradiction class, which more often shifts to neutral than entailment when negation is added. In the ZH data we also see a slightly different pattern than for the rest with respect to the change of the entailment label. It appears that negation is important across examples with this label at a lower rate for these subsets, i.e. the label remains unchanged more often between M^- and $M_O^{\pm}$ (at a rate of 0.46).

5 Probing Negation-Awareness

We now use our new multilingual benchmark to probe the negation awareness of a multilingual language model fine-tuned for the NLI task.

Multilingual language model for NLI We fine-tune the cased version of mBERT on the English MNLI training data using the standard sequence classification approach for sentence pairs (Devlin et al., 2019) and follow Hossain et al. (2020b) in training the model for 3 epochs, with a batch size of 32 and a learning rate of 2e-5. We then evaluate model performance on the minimal pairs. For the languages other than EN, the evaluation is considered a zero-shot setup, as the model has not seen any NLI training data in the target language.

Performance on minimal pairs Our benchmark is designed to draw conclusions about a model's reaction to the phenomenon of negation, by evaluating model predictions on both parts of the minimal pair.[5] To understand if and under which circumstances a model adequately processes negation, we are interested in cases where the model seemingly correctly processes a negation, i.e. it correctly labels an example containing a negation cue, but mispredicts once the negation is removed. We quantify this behaviour as the percentage change between the correct predictions on the original example (M^-), and the correct predictions on the original example *and* the modified example ($M^- \bowtie M_O^{\pm}$). While we want the model to perform well on all minimal pairs, considering minimal pairs that differ in important and unimportant negations separately allows us to gain better insights into model behaviour. The main results that we discuss in the following are shown in Figure 4.

[5] Performance results on the newly created examples in isolation are provided in Appendix C.

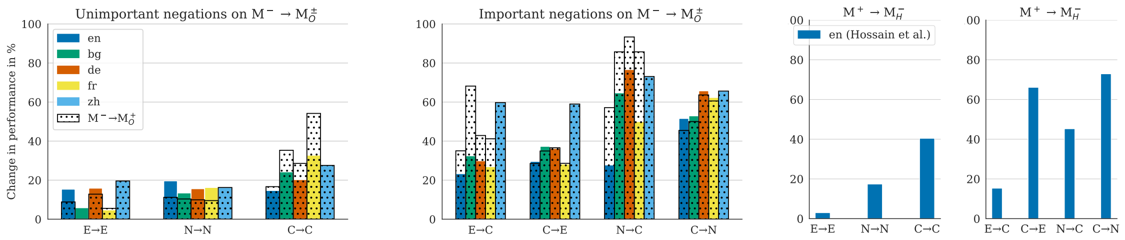


Figure 4: Performance loss on unimportant (left part) and important (right part) negations. The filled bars refer to performance on the full set of minimal pairs $M^- \bowtie M_O^\pm$. The dotted bars refer to performance on the subset of minimal pairs that does not contain any negation cue in the modified example ($M^- \bowtie M_O^+$). We do not include E→N and N→E on the x-axis, as we have very few pairs for this category (< 5 pairs).

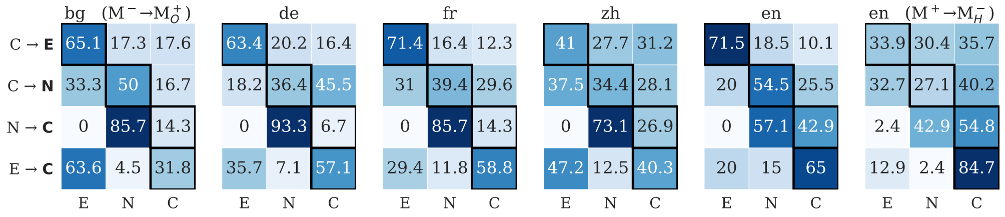


Figure 5: Confusions on minimal pairs with important negation. We show model predictions on the modified example, given that the corresponding original example was predicted correctly, with gold labels on the y-axis and predicted labels on the x-axis. The black frames indicate the percentage of correctly handled minimal pairs. The first five plots show results on the subset of our minimal pairs without negation mismatches ($M^- \bowtie M_O^+$); the last plot is on the minimal pairs derived from Hossain et al. (2020b) ($M^+ \bowtie M_H^-$).

Unimportant negations Performance on minimal pairs with unimportant negations can reveal if the model uses negations cues as shallow statistical cues without correctly modelling their effect on sentence semantics. Here, the model should stick with its prediction when the negation is removed. If the model changes its prediction, the initial correct prediction was due to the presence of the negation cue in the data (even though the negation cue is irrelevant for inferring the correct label). Possible explanations for the misprediction are (a) The model exploits the negation as a statistical cue without modeling its effect on sentence semantics; (b) The model incorrectly considers the negation to be relevant for inference.

Performance drops on minimal pairs with unimportant negations are shown on the left of Figure 4. We first focus on the filled bars that indicate performance on the complete set of minimal pairs $M^- \bowtie M_O^\pm$. We observe the largest performance drop for the contradiction class, across all languages (except EN), indicating that the model incorrectly changes its prediction after negation removal more often than for entailment or neutral. This is in line with the previously reported bias of negation cues indicating contradiction (Gururangan et al., 2018; Naik et al., 2018). Interestingly, in contrast to the other languages, for EN the largest performance drop is observed for the neutral class. The confusion matrices for this experiment (see Appendix D) show a high number of FPs for the contradiction class, consistently across all languages, and most pronounced for EN. This seems counter-intuitive, but might be in line with the *negation mismatch bias* (Dasgupta et al., 2018; McCoy and Linzen, 2019): Recall that in case an example in M^- contains a negation cue in premise *and* hypothesis, its counterpart in $M_O^\pm$ contains a negation cue in either premise *or* hypothesis, i.e. a negation mismatch. This applies to around 29% of the modified examples across languages.

Confirming negation mismatch bias To confirm this bias, we re-compute results on the subset of minimal pairs that does not contain any negation cues in the modified example ($M^- \bowtie M_O^+$). Results are shown in Figure 4 as dotted bars. We see that the performance drop for C → C increases for all languages, indicating that the model makes more mistakes on the minimal pairs when it cannot rely on the negation mismatch in the data. This confirms the previously reported bias for EN: negation mismatch makes a model predict contradiction, and by removing examples with negation mismatch we make it harder for the model to predict this class.

Important negations Performance on minimal pairs that differ in important negations can reveal if a model completely ignores negation cues, and to what extent it can model the effect of negation on sentence semantics (and its implication for inference). If the model mispredicts a modified example, it either (a) ignores negation completely and always makes the same prediction regardless of presence or absence of the negation cue, (b) incorrectly considers the negation to be irrelevant, or (c) changes to an incorrect prediction, either because it incorrectly models how the negation affects sentence semantics, or because it does not correctly model the inference. In the right part of Figure 4, we see that results are also affected by the negation mismatch bias: the performance drop increases for minimal pairs where the modified example is a →C after the pairs in line with the bias are removed. Further, we see that E → C and C → E switches are easier for the model than N → C and C → N across all languages. This also holds for ZH, however here the drop on the former two categories is much higher than for any of the other languages. Other than ZH, we observe the highest performance drops for DE and BG on the N → C category. In the confusion matrices in Figure 5, we see that for most languages (except DE and EN), for C → N, the largest source of error is a misprediction as E, i.e. the model flips its prediction, but to a non-neutral class.

Premise-hypothesis similarity One possible explanation for why predicting the modified examples in the C → N category is challenging might be a high content overlap in these modified `neutral` examples, as high content overlap has been reported to be indicative of the non-neutral categories (Dasgupta et al., 2018; Naik et al., 2018; McCoy et al., 2019). When deriving new examples from E or C examples, we can expect these newly created `neutral` examples to have a degree of content overlap that is closer to the degree of content overlap observed in original E and C examples. The premise-hypothesis cosine similarities[6] in Table 6 show that this holds for EN, BG and ZH, indicating why a model that strongly relies on low content overlap to identify `neutral` examples might fail on the minimal pairs in the C → N category.

lang	E	N	C	N → N	C → N
EN	0.917	0.913	0.915	0.928	0.939
BG	0.947	0.946	0.948	0.944	0.967
DE	0.951	0.949	0.954	0.951	0.950
FR	0.962	0.957	0.961	0.960	0.958
ZH	0.880	0.878	0.883	0.880	0.892

Table 6: Average premise-hypothesis cosine similarities for original examples and examples that became `neutral` after modification. For EN, BG, and ZH, similarity for C → N is higher than original C.

Comparison EN and zero-shot transfer Comparing EN with the zero-shot transfer languages on important negations, we observe the largest difference for the N → C category, that seems to be particularly difficult in the transfer setup. The performance drop for EN is consistently lower than for the zero-shot setup (except for FR on C → E). On the unimportant negations, for `contradiction` examples, the performance drop in the zero-shot setup exceeds EN by far, even more so when the negation mismatch bias cannot be exploited. This might indicate that in the zero-shot setup, the model relies on spurious cues even more than for predicting EN examples. For `neutral`, the performance drop in the zero-shot experiments is almost even across the languages. For `entailment` examples, we observe the largest variation between languages, with BG correctly solving all minimal pairs after the examples with negation mismatch are removed.

Insights from removing versus adding negation Comparing the above discussed patterns for predictions on our minimal pairs M_O with those on the minimal pairs M_H, we find that the most challenging categories are complementary: on the latter pairs, the most challenging categories are C → E and C → N (see the right-most subplot in Figure 4), i.e. examples that originally were contradictions and changed the label after adding a negation. This indicates that insights from both types of minimal pairs can complement each other.

6 Conclusion

In this work, we derived a multilingual benchmark for testing negation awareness using minimal pairs, and demonstrated its use to gain insights into a model's prediction behaviour. Using our benchmark, we confirm that biases previously reported

[6] Computed on their respective [CLS] representations

for English do transfer in a zero-shot transfer setup, and that the problem of exploiting shallow statistical cues might be larger in the latter setup.

7 Acknowledgements

Mareike Hartmann was funded by the Lundbeck foundation in the context of the BrainDrugs project and the German Federal Ministry of Research (BMBF) in the context of the XAINES project. Miryam de Lhoneux was funded by the Swedish Research Council (grant 2020-00437).

References

Carlos Aspillaga, Andrés Carvallo, and Vladimir Araujo. 2020. Stress test evaluation of transformer-based models in natural language understanding tasks. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1882–1894, Marseille, France. European Language Resources Association.

Jeremy Barnes, Lilja Øvrelid, and Erik Velldal. 2019. Sentiment analysis is not solved! assessing and probing sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 12–23, Florence, Italy. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Federico Fancellu and Bonnie Webber. 2015. Translating negation: A manual error analysis. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 2–11.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online. Association for Computational Linguistics.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

F. H. Heinemann. 2015. VIII.—The Meaning of Negation. *Proceedings of the Aristotelian Society*, 44(1):127–152.

Laurence Horn. 1989. A natural history of negation.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. 2020a. It's not a non-issue: Negation as a source of error in machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3869–3885, Online. Association for Computational Linguistics.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020b. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Devin Johnson, Denise Mak, Andrew Barker, and Lexi Loessberg-Zahl. 2020. Probing for multilingual numerical understanding in transformer-based language models. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 184–192, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Bingzhi Li and Guillaume Wisniewski. 2021. Are neural networks extracting linguistic properties or memorizing training data? an observation with a multilingual probe for predicting tense. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3080–3089, Online. Association for Computational Linguistics.

Shoushan Li and Chu-Ren Huang. 2009. Sentiment classification considering negation and contrast transition. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 307–316, Hong Kong. City University of Hong Kong.

Qianchu Liu, Federico Fancellu, and Bonnie Webber. 2018. Negpar: A parallel corpus annotated for negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Richard T McCoy and Tal Linzen. 2019. Non-entailed subsequences as a challenge for natural language inference. *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 358–360.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Roser Morante and Walter Daelemans. 2012. Conandoyle-neg: Annotation of negation cues and their scope in conan doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1563–1568.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Vinit Ravishankar, Memduh Gökırmak, Lilja Øvrelid, and Erik Velldal. 2019a. Multilingual probing of deep pre-trained contextual encoders. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 37–47, Turku, Finland. Linköping University Electronic Press.

Vinit Ravishankar, Lilja Øvrelid, and Erik Velldal. 2019b. Probing multilingual sentence representations with X-probe. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 156–168, Florence, Italy. Association for Computational Linguistics.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8713–8721.

Ieva Staliūnaitė and Ignacio Iacobacci. 2020. Compositional and lexical semantics in RoBERTa, BERT and DistilBERT: A case study on CoQA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7046–7056, Online. Association for Computational Linguistics.

Rosalind Thornton and Graciela Tesan. 2013. Sentential negation in early child english. *Journal of Linguistics*, 49(2):367–411.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Xiaodan Zhu, Hongyu Guo, Saif Mohammad, and Svetlana Kiritchenko. 2014. An empirical study on the effect of negation words on sentiment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 304–313, Baltimore, Maryland. Association for Computational Linguistics.

A Issues with Translation Quality in the XNLI Datasets

In the process of modifying the XNLI data, we came across numerous instances of incorrectly labeled examples in the non-English examples. In some cases, the fact that premise and hypothesis were translated independently (see (Conneau et al., 2018) for a description of the translation process) resulted in a grammatical or semantic mismatch.

Grammatical mismatch In pair 748 in the Bulgarian data, for example, we see a gender mismatch between premise and hypothesis, that can be translated as: ... I wasn't even happy[FEM] ||| I was so happy[MASC]. The `contradiction` label is appropriate for the EN version (*P: I wasn't even happy, H: I was happy*), but not for the BG examples considering that the premise and the hypothesis could not refer to the same person.

Poor translation quality In other cases, the translation of the English examples is just incorrect, which raises doubts about the quality of the workforce hired to do it. In pair 7033 in the Bulgarian data, for example, "Blood and flood are not like food" is translated to "Blood and flood are not a joke", and in pair 6922 'break' is translated into 'break up' instead.

In all of the cases discussed above, the confusion in the translation makes the label of the translated example *neutral*, as premise and hypothesis become unrelated. In the following example, the incorrect translation instead leads to a change from *contradiction* to *entailment*: the premise in "Others answered the question , but Keyes stuffed it . ||| Keyes didn 't answer the question ."' is translated in Bulgarian to "Others answered the question , but Keyes avoided it ."

B Distribution of Negation Cues

The ten most frequent negation cues per language are listed in Table 7.

C Performance on Newly Created Examples

In addition to reporting performance on minimal pairs, we measure the difficulty of our newly created examples $\mathrm{M}_O^{\pm}$ by comparing performance to the performance on the original XNLI data D. Figure 6 shows the results averaged over 3 model runs. Note that as the minimal pairs derived from Hossain et al. (2020b) are based on MNLI *training* examples, we remove those from the training data before training the model for the EN experiment on their data. Focusing on our EN and BG datasets, as expected, we see a general drop in performance between EN and the other languages, as for those we do zero-shot predictions. We see that the performance on the original XNLI data without any negation markers (D^+), and the performance on the subset of the XNLI with negation markers that is the base of our minimal pairs (M^-), varies strongly by class: in M^-, we see that performance on the contradiction class is high. This might be explained by the *negation cue indicates contradiction* bias. The pattern holds across all languages (weakest for ZH), indicating that the underlying cause transfers across languages. Comparing $\mathrm{M}_O^{\pm}$ and D^+ reveals if the newly created examples are more difficult to predict than the original examples without negation markers, and indeed they are, with the largest drop again being observed for ZH. Performance drops the most for the neutral class, this again holds across all languages. For the minimal pairs created from the (Hossain et al., 2020b) dataset, we compare their modified examples M_H^{+} to their counterparts in the original data M^+, and here we observe an even larger drop in performance on the new examples.

D Performance on Minimal Pairs

We present confusion matrices for predictions on the full set of minimal pairs $\mathrm{M}^- \bowtie \mathrm{M}_O^{\pm}$ in Figure 10. For the unimportant negations, we observe a high number of FPs for the contradiction class, consistently across all languages, and most pronounced for EN. Total performance (in percent of correctly labeled pairs) along with the change in performance for our minimal pairs can be found in Figures 8 (unimportant negations) and 9 (important negations). Results for the minimal pairs derived from (Hossain et al., 2020b) are shown in Figure 7.

Rank		EN		DE		BG		FR		ZH
1	685	not	526	nicht	547	не	260	ne	480	不
2	127	no	216	keine	63	никога не	157	pas	21	没
3	126	never	89	nie	47	няма	72	jamais	21	未
4	10	nobody	59	nichts	25	няма да	42	n'est pas	15	没有
5	9	without	50	keinen	23	никога	30	aucun	12	从来没有
6	9	unable	30	niemand	17	никой не	29	ne sont pas	10	不要
7	8	refused	30	kein	15	нямаше	16	rien	9	从不
8	5	no way	22	niemals	12	изобщо не	13	plus	9	并未
9	5	either	20	ohne	8	никакви	11	ne peut pas	7	无
10	5	nowhere	15	noch nie	8	нямат	10	non	6	不用

Table 7: The ten most frequent negation cues in the final set of minimal pairs.

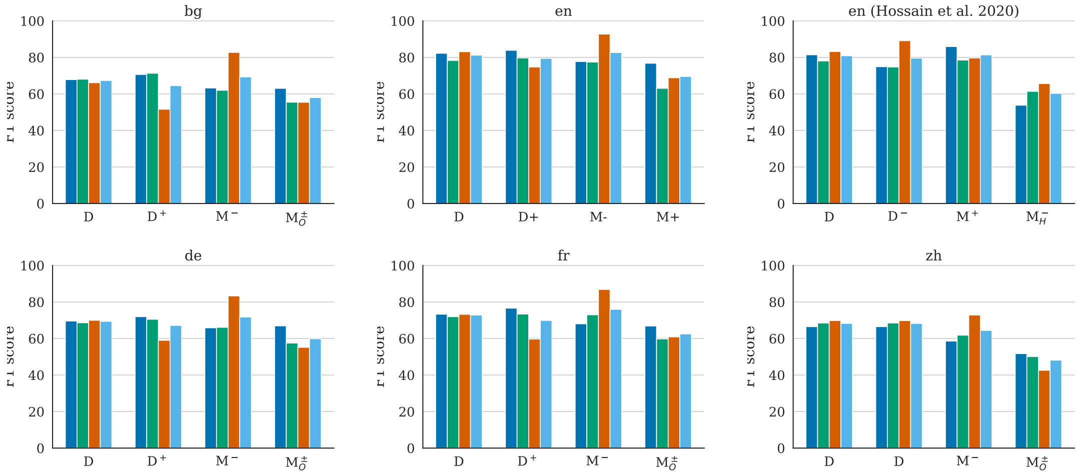


Figure 6: Comparison of model predictions on subsets of the original XNLI data (D) that do not contain any negation markers (D^+), and our newly created modified examples $M_O^{\pm}$. M^- is the subset of the original data that forms the base for our minimal pairs, i.e. examples that contain one negation cue in premise and/or hypothesis. For the Hossain et al. (2020b) dataset, where original examples are modified by adding negation cues, the modified examples are indicated as M_H^+.

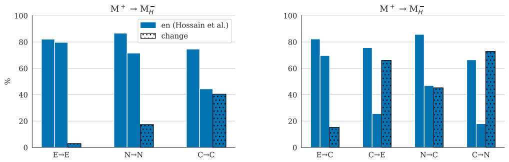


Figure 7: Absolute performance (full bars) and percentage change in performance (dotted bars) on the minimal pairs derived from (Hossain et al., 2020b). The first full bar indicates the fraction of correctly predicted examples in M^+, and the second full bar indicates for how many of these examples the corresponding modified example in M_H^+ is predicted correctly as well. The dotted bar indicates the percentage change between both values.

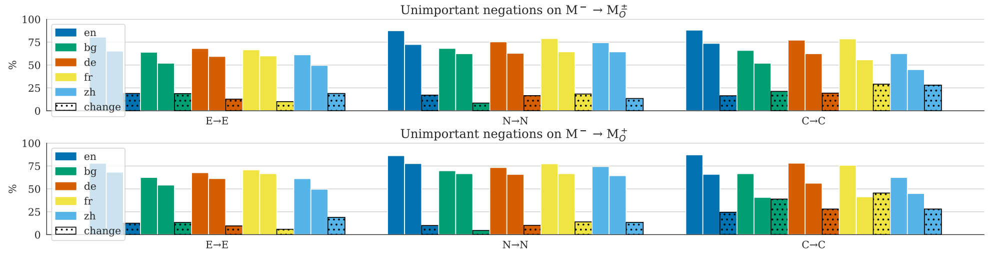


Figure 8: Absolute performance (full bars) and percentage change in performance (dotted bars) on unimportant negations, on the full set of minimal pairs (upper part) or the subset without any negation mismatches (lower part). The first full bar indicates the fraction of correctly predicted examples in M^-, and the second full bar indicates for how many of these examples the corresponding modified example in $M_O^\pm$ is predicted correctly as well. The dotted bar indicates the percentage change between both values.

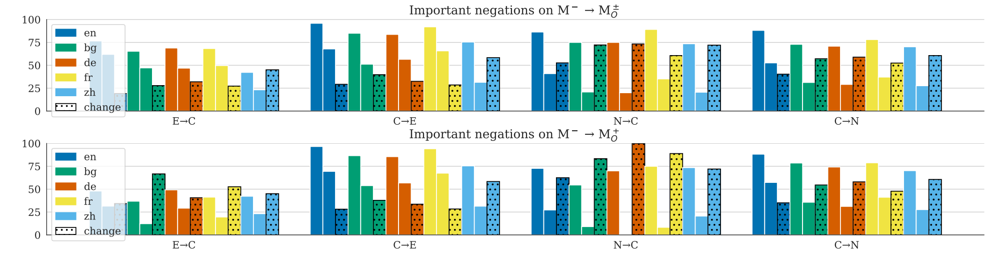


Figure 9: Absolute performance (full bars) and percentage change in performance (dotted bars) on important negations, on the full set of minimal pairs (upper part) or the subset without any negation mismatches (lower part).

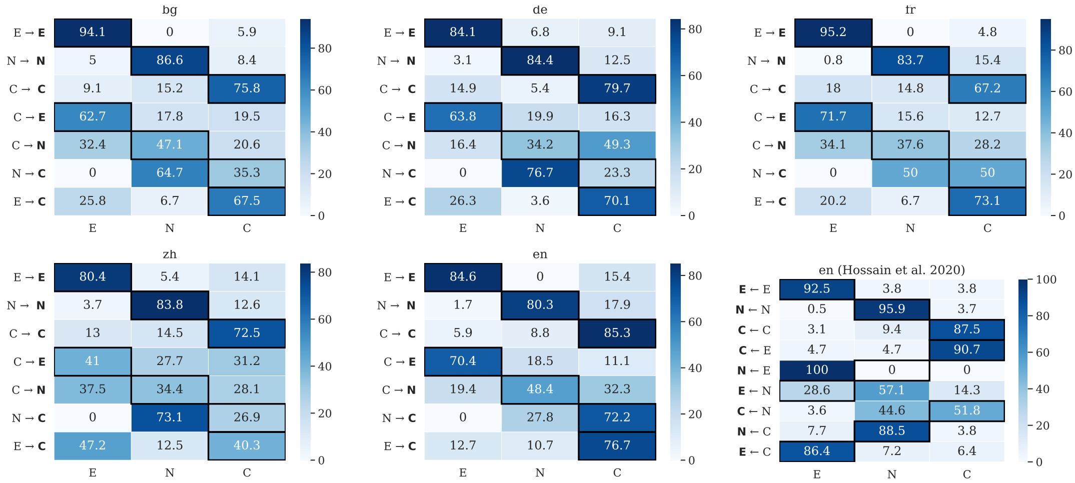


Figure 10: Confusion matrices for predictions on the full set of minimal pairs $M^- \bowtie M_O^\pm$, for unimportant negations (first three rows) and important negations (last 4 rows). We show model predictions on the modified example, given that the corresponding original example was predicted correctly, with gold labels on the y-axis and predicted labels on the x-axis. The black frames indicate the percentage of correctly handled minimal pairs.