HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

UNIVERSITY OF GOTHENBURG

CLASP centre for linguistic theory and studies in probability

NLI Data Sanity Check

Assessing the Effect of Data Corruption on Model Performance

Aarne Talman*, Marianna Apidianaki*, Stergios Chatzikyriakidis and Jörg Tiedemann***

** Department of Digital Humanities, University of Helsinki*
*** CLASP, Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg*

aarne.talman@helsinki.fi

Introduction

Natural Language Inference (NLI) is the problem of determining whether a sentence (hypothesis) can be inferred from a another sentence (premise).

A simple example:

- Premise: *A group of people are standing in front of a building.*
- Hypothesis: *A group of people are in front of a building.*

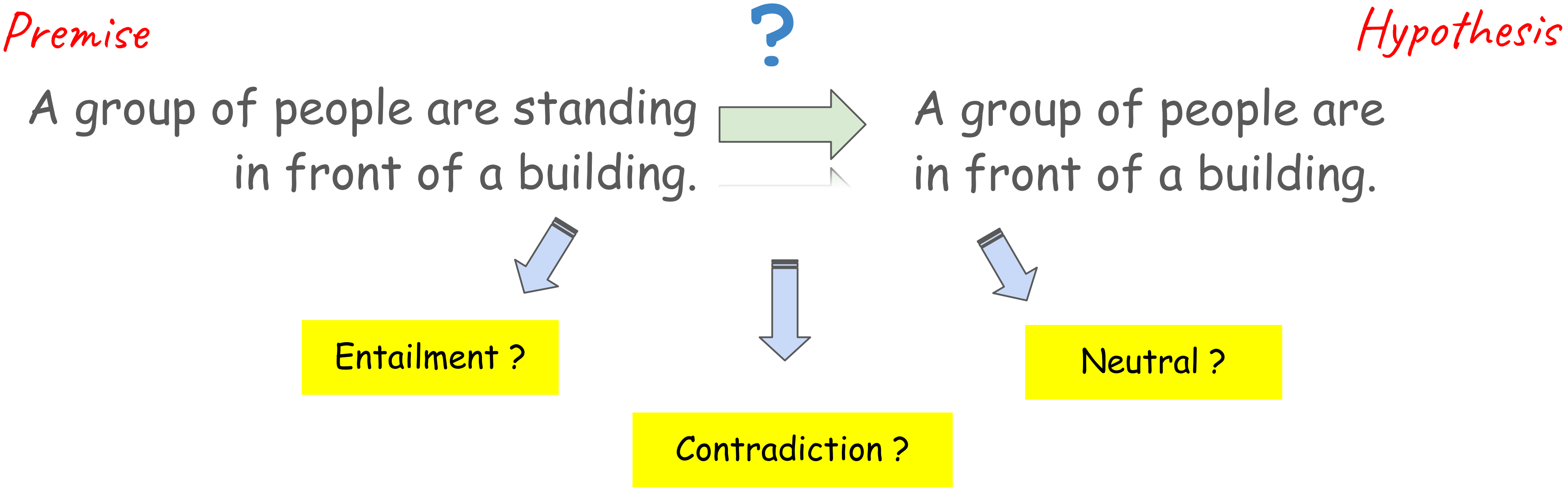
A typical NLI task involves classification of such hypothesis-premise pairs into entailments, contradictions or neutral.

Popular NLI datasets: SNLI (Bowman et al. 2015), MNLI (Williams et al. 2018)

Natural Language Inference

Datasets

- SNLI (Bowman et al., 2015)
- MNLI (Williams et al., 2018)
- ANLI (Nie et al, 2020)



Premise

A group of people are standing in front of a building.

?

Hypothesis

A group of people are in front of a building.

Entailment ?

Contradiction ?

Neutral ?

NLI Benchmarks

Serve as testbeds for measuring the models’ language understanding capabilities…

… based on the assumption that models should understand, or at least somehow encode the meaning of the processed sentences, in order to determine whether they entail each other

But, is it really the case?

Known problems with NLI Datasets

NLI datasets have been shown to contain annotation artefacts and other statistical biases.

Examples:

- Good performance in NLI tasks using only the hypothesis sentence as input (e.g. Poliak et al. 2018 and Gururangan et al. 2018).
- Notably, 90% of the hypotheses that denote a contradiction in the SNLI dataset contain the verb sleep and its variants (sleeping, asleep).
- Difficulties in generalisation across benchmark tasks: Talman & Chatzikyriakidis (2019) show that models trained on data drawn from one NLI benchmark (e.g. SNLI) do not perform well when tested on data from another benchmark (e.g. MNLI).
- Recently Pham et al. (2020) have shown that high accuracy can be achieved even after shuffling the word order of the NLI sentences.

Annotation artefacts and statistical biases

Hypothesis-only testing

(Poliak et al., 2018; Gururangan et al., 2018)

Contradictions marked with negation, entailments with generic nouns

(Lai and Hockenmaier, 2014; Marelli et al., 2014; Gururangan et al., 2018)

90% of contradiction hypotheses in SNLI contain variants of *sleep* (e.g., *sleeping*, *asleep*)

(Poliak et al., 2018; Gururangan et al., 2018)

Lack of generalisation across benchmarks (e.g., SNLI => MNLI)

(Talman and Chatzikyriakidis, 2019)

High accuracy after word shuffling in NLI sentences

(Pham et al., 2020)

How do you determine if an NLI dataset is of good quality?

How to determine if an NLI dataset is high quality?

Approach

Arguably as NLI datasets test semantic relationships between sentences, models should *understand* or at least somehow *encode the meaning* of the sentences and then determine whether they entail each other or are in contradiction.

Our approach: **try to corrupt NLI sentences in a systematic way and test what is the impact on model performance:**

- Corrupt datasets by removing words belonging to a specific word class, e.g. **verbs** or **nouns**, to create sentences that don't really make sense any more.
- If model accuracy on the corrupted data remains high, then the dataset is likely to contain statistical biases and artefacts that guide prediction.
- Inversely, a large decrease in model accuracy indicates that the original dataset provides a proper challenge to the models' reasoning capabilities.

Our Data Sanity Check Approach

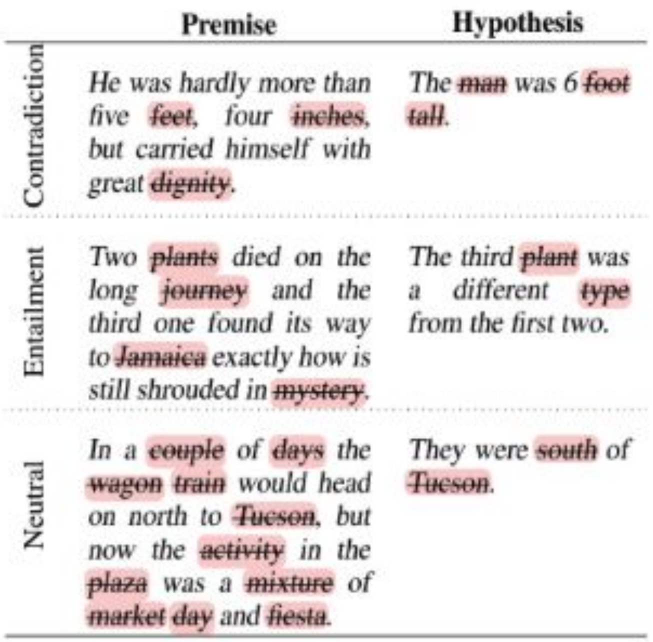
Systematically corrupt NLI sentences

- ○ Remove words of a specific grammatical category (e.g., verbs, nouns, adjectives, adverbs)
- ○ Create sentences that often do not make sense!

Test impact on model performance

- ○ High model accuracy: the dataset is likely to contain statistical biases and artefacts that guide prediction.
- ○ Large decrease in model accuracy: the original dataset provides a proper challenge to the models' reasoning capabilities.

Train - NOUNS



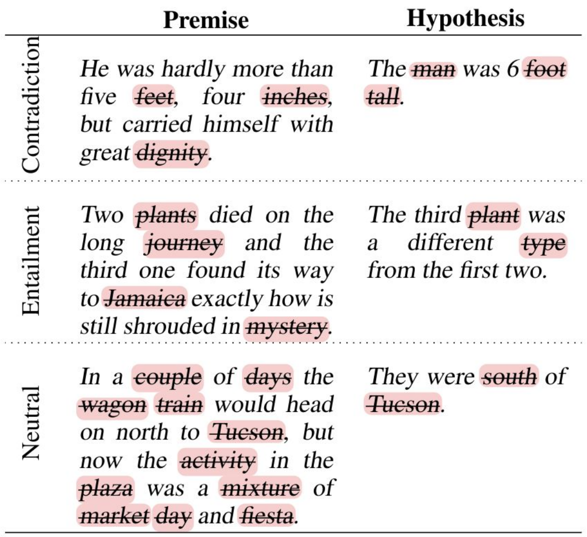
	Premise	Hypothesis
Contradiction	He was hardly more than five ~~feet~~, four ~~inches~~, but carried himself with great ~~dignity~~.	The ~~man~~ was 6 ~~foot~~ ~~tall~~.
Entailment	Two ~~plants~~ died on the long ~~journey~~ and the third one found its way to ~~Jamaica~~ exactly how is still shrouded in ~~mystery~~.	The third ~~plant~~ was a different ~~type~~ from the first two.
Neutral	In a ~~couple~~ of ~~days~~ the ~~wagon~~ ~~train~~ would head on north to ~~Tucson~~, but now the ~~activity~~ in the ~~plaza~~ was a ~~mixture~~ of ~~market~~ ~~day~~ and ~~fiesta~~.	They were ~~south~~ of ~~Tucson~~.

Datasets - MNLI (Williams et al. 2018)

We created 42 different configurations:

- 9 corruptions with specific word class(es) removed from MNLI: `-NUM, -CONJ, -ADV, -PRON, -ADJ, -DET, -VERB, -NOUN, -NOUN-PRON`
- 5 corruptions where only specific word classes are present (others removed): `NOUN+PRON+VERB, NOUN+ADJ+VERB, NOUN+VERB, NOUN+VERB+ADJ, NOUN+VERB+ADJ+ADV`
- 3 different experimental setups per corruption:
  - **Corrupt-Train:** only data in the training set has been corrupted
  - **Corrupt-Test:** only data in the test set has been corrupted*
  - **Corrupt-Train and Test:** both train and test sets have been corrupted

Examples:



	Premise	**Hypothesis**
Contradiction	*He was hardly more than five ~~feet~~, four ~~inches~~, but carried himself with great ~~dignity~~.*	*The ~~man~~ was 6 ~~foot~~ ~~tall~~.*
Entailment	*Two ~~plants~~ died on the long ~~journey~~ and the third one found its way to ~~Jamaica~~ exactly how is still shrouded in ~~mystery~~.*	*The third ~~plant~~ was a different ~~type~~ from the first two.*
Neutral	*In a ~~couple~~ of ~~days~~ the ~~wagon~~ ~~train~~ would head on north to ~~Tucson~~, but now the ~~activity~~ in the ~~plaza~~ was a ~~mixture~~ of ~~market~~ ~~day~~ and ~~fiesta~~.*	*They were ~~south~~ of ~~Tucson~~.*

Table 1: Sentence pairs from a corrupted MNLI training dataset where nouns have been removed.

* We use MNLI-matched development set for testing as annotated test sets are not publicly available

Corrupting MNLI (Williams et al. 2018)

42 different configurations

- 9 with specific word class(es) removed
  - `-NUM, -CONJ, -ADV, -PRON, -ADJ, -DET, -VERB, -NOUN, -NOUN-PRON`
- 5 with specific word classes present (others removed)
  - `NOUN+PRON+VERB, NOUN+ADJ+VERB, NOUN+VERB, NOUN+VERB+ADJ, NOUN+VERB+ADJ+ADV`
- 3 experimental setups per corruption
  - **Corrupt-Train:** corrupting the training set
  - **Corrupt-Test:** corrupting the test set (i.e. the MNLI-matched dev set)
  - **Corrupt-Train and Test:** corrupting both sets

Datasets - ANLI (Nie et al. 2018)

The Adversarial NLI benchmark (ANLI) was specifically designed to address some of the shortcomings of the previous NLI datasets.

We created 27 different configurations for ANLI:

- 8 corruptions with specific word class(es) removed from MNLI: `-NUM, -CONJ, -ADV, -PRON, -ADJ, -DET, -VERB, -NOUN`
- ANLI contains 3 datasets (rounds), R1, R2 and R3. Each dataset was collected using a human-and-model-in-the-loop approach, and they progressively increase in difficulty and complexity.
- For ANLI we only used the **Corrupt-Test** configuration.

Corrupting ANLI (Nie et al. 2018)

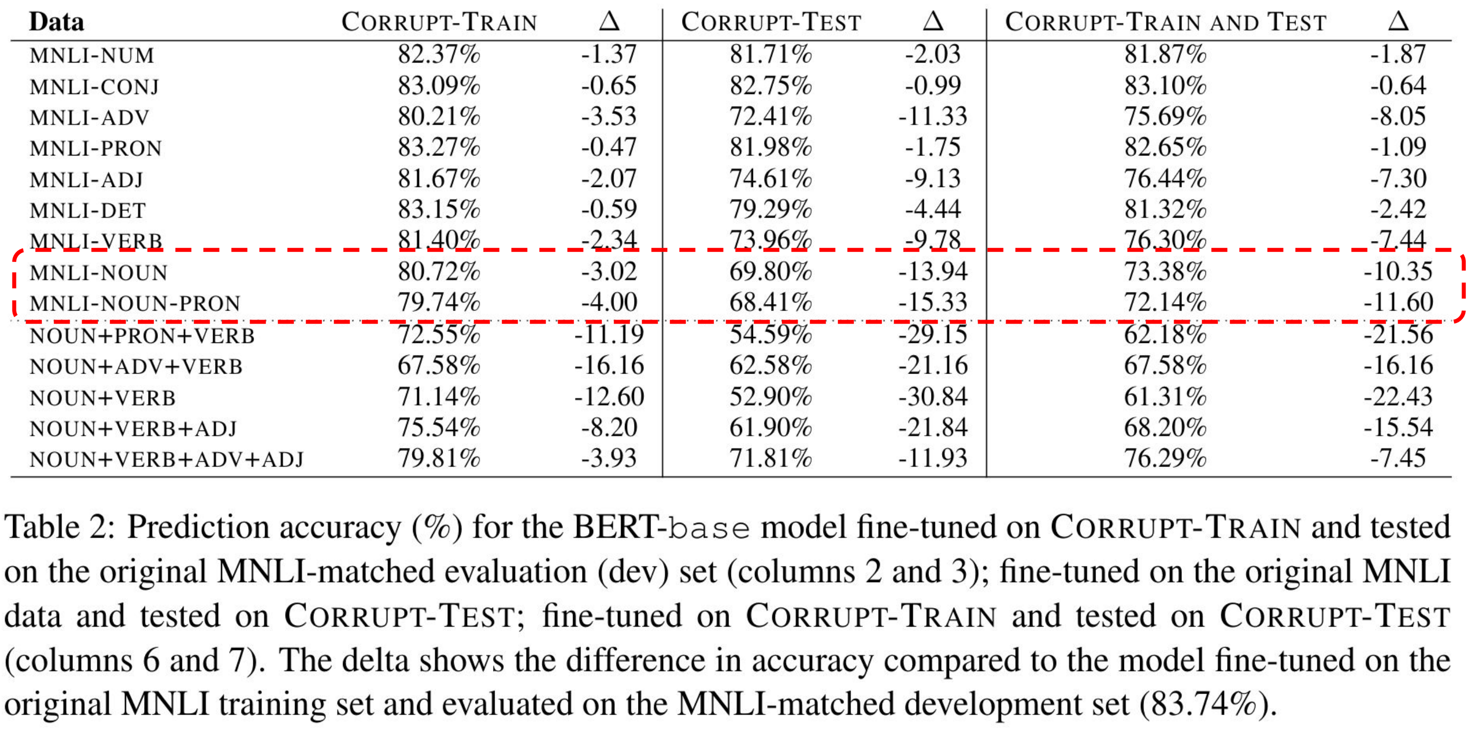
- The Adversarial NLI benchmark (ANLI) was specifically designed to address shortcomings of the previous NLI datasets.
- ANLI contains 3 datasets (rounds): R1, R2 and R3.
- Each dataset was collected using a human-and-model-in-the-loop approach, and they progressively increase in difficulty and complexity.

27 different configurations

- 8 corruptions with specific word class(es) removed from MNLI
  - `-NUM, -CONJ, -ADV, -PRON, -ADJ, -DET, -VERB, -NOUN`
- For ANLI we only used he **Corrupt-Test** experimental setup.

Results with corrupt MNLI and BERT (Devlin et al. 2018)

We use training and evaluation scripts provided by Google using the default hyperparameter values and other settings (`https://github.com/google-research/bert`)



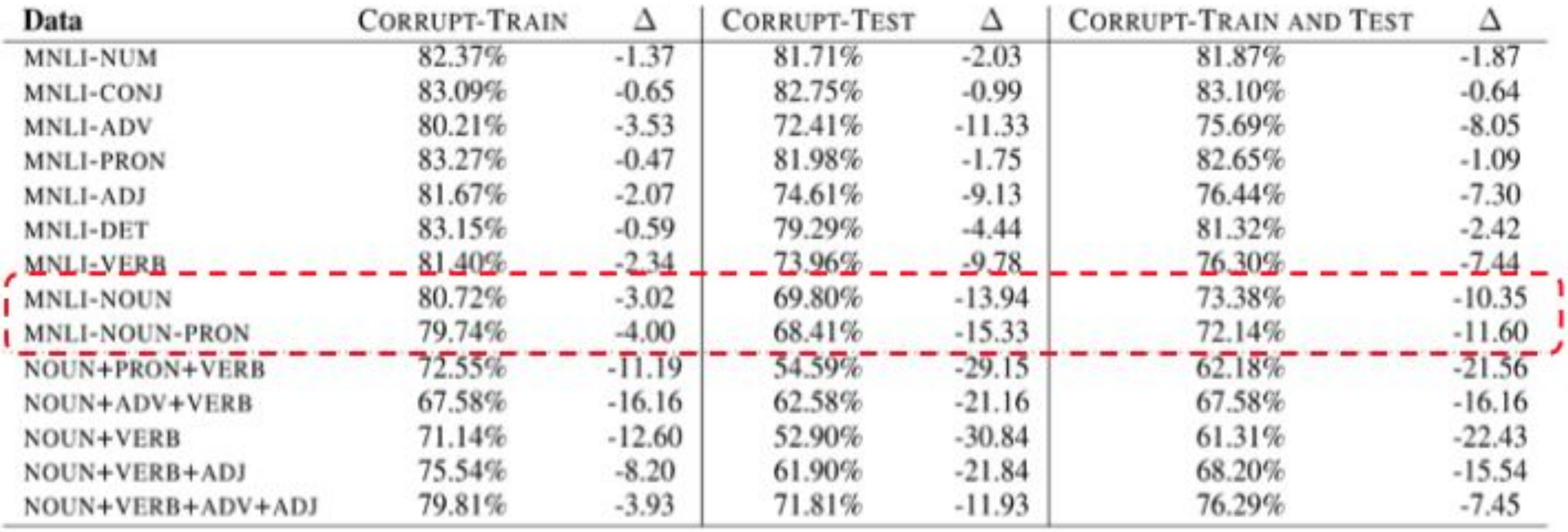
Data	CORRUPT-TRAIN	Δ	CORRUPT-TEST	Δ	CORRUPT-TRAIN AND TEST	Δ
MNLI-NUM	82.37%	-1.37	81.71%	-2.03	81.87%	-1.87
MNLI-CONJ	83.09%	-0.65	82.75%	-0.99	83.10%	-0.64
MNLI-ADV	80.21%	-3.53	72.41%	-11.33	75.69%	-8.05
MNLI-PRON	83.27%	-0.47	81.98%	-1.75	82.65%	-1.09
MNLI-ADJ	81.67%	-2.07	74.61%	-9.13	76.44%	-7.30
MNLI-DET	83.15%	-0.59	79.29%	-4.44	81.32%	-2.42
MNLI-VERB	81.40%	-2.34	73.96%	-9.78	76.30%	-7.44
MNLI-NOUN	80.72%	-3.02	69.80%	-13.94	73.38%	-10.35
MNLI-NOUN-PRON	79.74%	-4.00	68.41%	-15.33	72.14%	-11.60
NOUN+PRON+VERB	72.55%	-11.19	54.59%	-29.15	62.18%	-21.56
NOUN+ADV+VERB	67.58%	-16.16	62.58%	-21.16	67.58%	-16.16
NOUN+VERB	71.14%	-12.60	52.90%	-30.84	61.31%	-22.43
NOUN+VERB+ADJ	75.54%	-8.20	61.90%	-21.84	68.20%	-15.54
NOUN+VERB+ADV+ADJ	79.81%	-3.93	71.81%	-11.93	76.29%	-7.45

Table 2: Prediction accuracy (%) for the BERT-`base` model fine-tuned on CORRUPT-TRAIN and tested on the original MNLI-matched evaluation (dev) set (columns 2 and 3); fine-tuned on the original MNLI data and tested on CORRUPT-TEST; fine-tuned on CORRUPT-TRAIN and tested on CORRUPT-TEST (columns 6 and 7). The delta shows the difference in accuracy compared to the model fine-tuned on the original MNLI training set and evaluated on the MNLI-matched development set (83.74%).

Experimental Setup for MNLI

- We use the BERT-`base` model (Devlin et al., 2018)
- Training and evaluation scripts provided by Google, with the default hyperparameter settings (`https://github.com/google-research/bert`)
- We measure the model's prediction accuracy when
  - fine-tuned on Corrupt-TRAIN and tested on the original MNLI-matched evaluation (dev) set
  - fine-tuned on the original MNLI data and tested on Corrupt-TEST
  - fine-tuned on Corrupt-TRAIN and tested on Corrupt-TEST

Results on MNLI

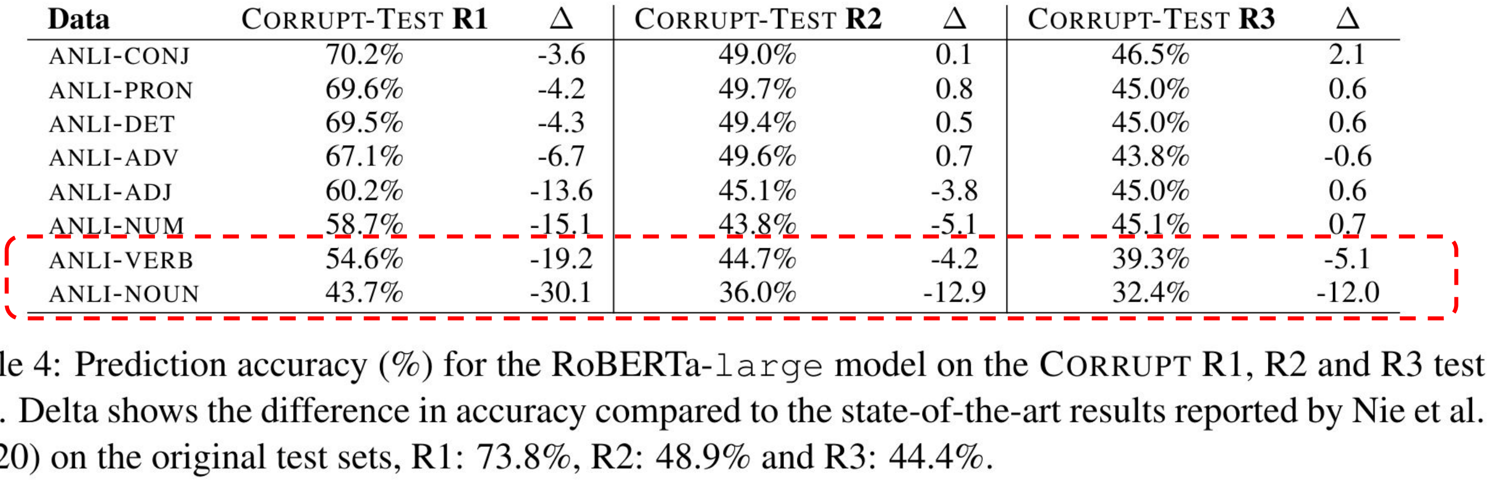


Data	Corrupt-Train	Δ	Corrupt-Test	Δ	Corrupt-Train and Test	Δ
MNLI-NUM	82.37%	-1.37	81.71%	-2.03	81.87%	-1.87
MNLI-CONJ	83.09%	-0.65	82.75%	-0.99	83.10%	-0.64
MNLI-ADV	80.21%	-3.53	72.41%	-11.33	75.69%	-8.05
MNLI-PRON	83.27%	-0.47	81.98%	-1.75	82.65%	-1.09
MNLI-ADJ	81.67%	-2.07	74.61%	-9.13	76.44%	-7.30
MNLI-DET	83.15%	-0.59	79.29%	-4.44	81.32%	-2.42
MNLI-VERB	81.40%	-2.34	73.96%	-9.78	76.30%	-7.44
MNLI-NOUN	80.72%	-3.02	69.80%	-13.94	73.38%	-10.35
MNLI-NOUN-PRON	79.74%	-4.00	68.41%	-15.33	72.14%	-11.60
NOUN+PRON+VERB	72.55%	-11.19	54.59%	-29.15	62.18%	-21.56
NOUN+ADV+VERB	67.58%	-16.16	62.58%	-21.16	67.58%	-16.16
NOUN+VERB	71.14%	-12.60	52.90%	-30.84	61.31%	-22.43
NOUN+VERB+ADJ	75.54%	-8.20	61.90%	-21.84	68.20%	-15.54
NOUN+VERB+ADV+ADJ	79.81%	-3.93	71.81%	-11.93	76.29%	-7.45

- Delta shows the difference in accuracy compared to the BERT--`base` model fine-tuned on the original MNLI training set and evaluated on the MNLI-matched development set (83.74%).

Results with corrupt ANLI and RoBERTa (Liu et al. 2019)

We use training and evaluation scripts provided by Liu et al. using the default hyperparameter values and other settings (`https://github.com/facebookresearch/anli`)



Data	CORRUPT-TEST **R1**	Δ	CORRUPT-TEST **R2**	Δ	CORRUPT-TEST **R3**	Δ
ANLI-CONJ	70.2%	-3.6	49.0%	0.1	46.5%	2.1
ANLI-PRON	69.6%	-4.2	49.7%	0.8	45.0%	0.6
ANLI-DET	69.5%	-4.3	49.4%	0.5	45.0%	0.6
ANLI-ADV	67.1%	-6.7	49.6%	0.7	43.8%	-0.6
ANLI-ADJ	60.2%	-13.6	45.1%	-3.8	45.0%	0.6
ANLI-NUM	58.7%	-15.1	43.8%	-5.1	45.1%	0.7
ANLI-VERB	54.6%	-19.2	44.7%	-4.2	39.3%	-5.1
ANLI-NOUN	43.7%	-30.1	36.0%	-12.9	32.4%	-12.0

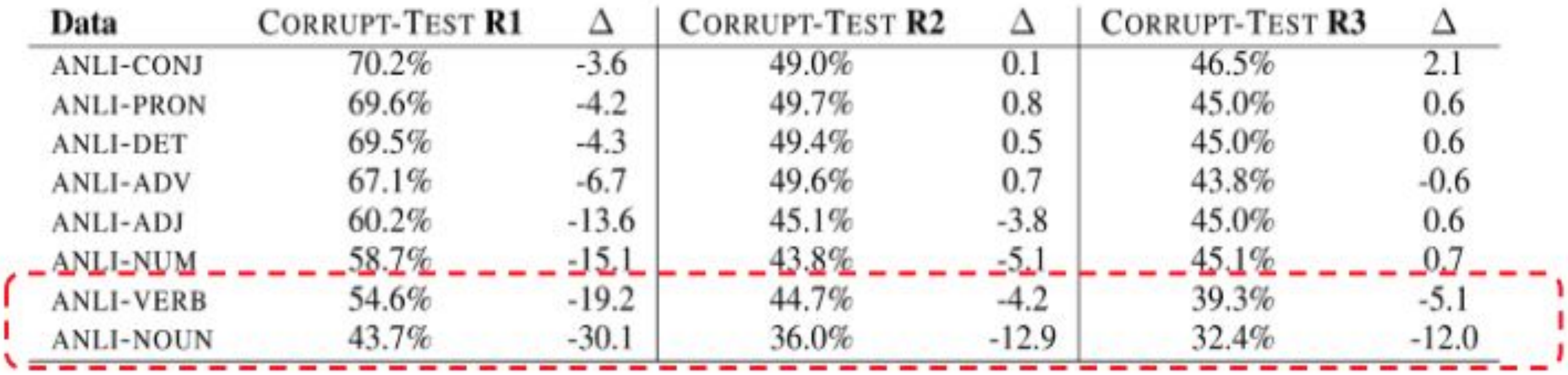
Table 4: Prediction accuracy (%) for the RoBERTa-`large` model on the CORRUPT R1, R2 and R3 test sets. Delta shows the difference in accuracy compared to the state-of-the-art results reported by Nie et al. (2020) on the original test sets, R1: 73.8%, R2: 48.9% and R3: 44.4%.

Experimental Setup for ANLI

- We use the RoBERTa-`large` model (Liu et al., 2019)
- Training and evaluation scripts provided by Liu et al. using the default hyperparameter values and other settings (`https://github.com/facebookresearch/anli`)

- Evaluation
  - We measure the prediction accuracy of RoBERTa-`large` on the Corrupt R1, R2 and R3 test sets

Results for ANLI



Data	Corrupt-Test **R1**	Δ	Corrupt-Test **R2**	Δ	Corrupt-Test **R3**	Δ
ANLI-CONJ	70.2%	-3.6	49.0%	0.1	46.5%	2.1
ANLI-PRON	69.6%	-4.2	49.7%	0.8	45.0%	0.6
ANLI-DET	69.5%	-4.3	49.4%	0.5	45.0%	0.6
ANLI-ADV	67.1%	-6.7	49.6%	0.7	43.8%	-0.6
ANLI-ADJ	60.2%	-13.6	45.1%	-3.8	45.0%	0.6
ANLI-NUM	58.7%	-15.1	43.8%	-5.1	45.1%	0.7
ANLI-VERB	54.6%	-19.2	44.7%	-4.2	39.3%	-5.1
ANLI-NOUN	43.7%	-30.1	36.0%	-12.9	32.4%	-12.0

- Delta shows the difference in accuracy compared to the state-of-the-art results reported by Nie et al. (2020) on the original test sets
    - R1: 73.8%
    - R2: 48.9%
    - R3: 44.4%

Discussion

- Our results confirm previous findings that neural network models are able to solve NLI tasks like MNLI by using statistical cues and artefacts in the data.
- Instead of learning to “understand” the sentences in NLI datasets, Transformer-based models like BERT and RoBERTa can utilise other factors from the datasets to guide predictions.
- Our method demonstrates the superior quality of the ANLI datasets which was specifically designed to get rid of annotation artefacts and biases in the data.

Findings

- Our results show a lower-than-expected decrease in performance for models fine-tuned/tested on corrupted data, where sentences are often unintelligible.

- They confirm that
  - neural network models are able to solve NLI tasks by relying on statistical cues and artefacts in the data
  - rather than "understanding" sentence meaning, Transformer-based models leverage other cues in the datasets to guide prediction.

- Our method demonstrates the superior quality of the ANLI datasets which was specifically designed to remove artefacts and biases from the data.

Future research

- Extending the proposed method to other natural language understanding (NLU) datasets and benchmarks.
- Our method can only indicate if an NLI/NLU dataset is of high or low quality, it does not reveal the actual biases in the dataset - this should be analysed in future research.
- Comprehensive NLU datasets evaluation methodology and design guidelines - what it takes to develop a good NLI/NLU dataset?

Future research

- Extending the proposed method to other natural language understanding (NLU) datasets and benchmarks.

- Our method can only indicate if a dataset is of high or low quality, it does not reveal the actual biases in the dataset - this should be analysed in future research.

- Comprehensive NLU datasets evaluation methodology and design guidelines - what it takes to develop a good NLI/NLU dataset?

Thank you!

aarne.talman@helsinki.fi

@AarneTalman