How Does Data Corruption Affect Natural Language Understanding Models?

Aarne Talman
University of Helsinki / Silo AI

@aarnetalman

Outline

1. Introduction and motivation
2. How does data corruption affect Natural Language Inference models?
3. What is the impact on the other Natural Language Understanding tasks?
4. Discussion and next steps

Motivation

Claim: Natural language understanding benchmark tasks (like GLUE) serve as testbeds for measuring models' language understanding capabilities

Assumption: Models should understand, or at least somehow encode the meaning of the processed sentences, in order to perform well in these tasks

Question: Is understanding really required for good performance?

Popular Natural Language Inference Datasets SNLI and MNLI are known to be problematic

Hypothesis-only testing

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

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

(Talman and Chatzikyriakidis, 2019)

Contradictions marked with negation, entailments with generic nouns

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

High accuracy after word shuffling in NLI sentences

(Pham et al., 2020)

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

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

Our approach

We corrupt sentences in the benchmark datasets 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 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.

Part 1

How does data corruption affect Natural Language Inference models?

How does data corruption affect Natural Language Inference models?

- Aarne Talman, Marianna Apidianaki, Stergios Chatzikyriakidis, Jörg Tiedemann. 2021. **NLI Data Sanity Check: Assessing the Effect of Data Corruption on Model Performance.** *Proceedings of NoDaLiDa*.

- Studied the impact on
  - MNLI - Multi-Genre Natural Language Inference (Williams at al. 2018)
  - ANLI - The Adversarial NLI benchmark (Nie et al. 2018)

	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 created 42 different configurations for the MNLI experiments

- 14 corruption types:
  - 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 type
  - 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

We created 27 different configurations for the ANLI experiments

- 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

- 3 datasets
- 8 corruption types per per dataset with specific word class(es) removed
  - -NUM, -CONJ, -ADV, -PRON, -ADJ, -DET, -VERB, -NOUN
- For ANLI we only used he Corrupt-Test experimental setup.

Experimental setup on 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

GitHub repo for our data: *https://github.com/Helsinki-NLP/nli-data-sanity-check*

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

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 on 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 on 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%.

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 some 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 also demonstrates the superior quality of the ANLI datasets which was specifically designed to remove artefacts and biases from the data.

Part 2

What is the impact on the other Natural Language Understanding tasks?

How Does Data Corruption Affect Natural Language Understanding Models?

- Aarne Talman, Marianna Apidianaki, Stergios Chatzikyriakidis, Jörg Tiedemann. 2022. **How Does Data Corruption Affect Natural Language Understanding Models? A Study on GLUE datasets.** *Proceedings of *SEM 2022.*

- Studied the impact on
  - linguistic acceptability (COLA)
  - paraphrasing (MRPC and QQP)
  - sentiment prediction (SST-2)
  - semantic textual similarity (STS-B)
  - natural language inference (QNLI, RTE, MNLI)

	Sentence 1	Sentence 2
paraphrase	~~Easynews~~ ~~Inc.~~ was subpoenaed late last ~~week~~ by the ~~FBI~~, which was seeking ~~account information~~ related to the ~~uploading~~ of the ~~virus~~ to the ~~ISP's~~ ~~Usenet~~ ~~news~~ ~~group~~ ~~server~~.	~~Easynews~~ ~~Inc.~~ said ~~Monday~~ that it was co-operating with the ~~FBI~~ in trying to locate the ~~person~~ who uploaded the ~~virus~~ to a ~~Usenet~~ ~~news~~ ~~group~~ hosted by the ~~ISP~~.
non-paraphrase	~~Arison~~ said ~~Mann~~ may have been one of the ~~pioneers~~ of the ~~world~~ ~~music~~ ~~movement~~ and he had a deep ~~love~~ of Brazilian ~~music~~.	~~Arison~~ said ~~Mann~~ was a ~~pioneer~~ of the ~~world~~ ~~music~~ ~~movement~~ – well before the ~~term~~ was coined – and he had a deep ~~love~~ of Brazilian ~~music~~.

Table 1: Example sentence pairs from the corrupted MRPC training dataset where all instances of nouns have been removed.

We created 192 different configurations for our GLUE experiments

- 8 GLUE tasks:
  - CoLa, MNLI, MRPC, QNLI, QQP, RTE, SST-2, STS-B
- 8 corruption types with specific word classes removed
  - ADJ, ADV, CONJ, DET, NOUN, NUM, PRON, VERB
- 3 experimental setups per corruption type and task
  - 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

Experimental setup on GLUE

- We fine-tuned a RoBERTa-base model in each of our 192 configurations. We use the same fine-tuning and evaluation set up for all the experiments.

- We measure the model's prediction accuracy when
  - fine-tuned on Corrupt-TRAIN and tested on the original test set
  - fine-tuned on the original MNLI data and tested on Corrupt-TEST
  - fine-tuned on Corrupt-TRAIN and tested on Corrupt-TEST

GitHub repo: *https://github.com/Helsinki-NLP/nlu-dataset-diagnostics*

Results

Results are reported as heat maps showing the delta to the baseline (non-corrupted) results using otherwise exactly the same experimental setup

Baseline results:

	Task	Baseline	Metric
COLA	The Corpus of Linguistic Acceptability (Warstadt et al., 2018)	64.05	Matthew's correlation
MNLI-M	Multi-Genre Natural Language Inference (Williams et al., 2018)	87.89	accuracy
MRPC	Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005)	88.73	accuracy
QNLI	Question Natural Language Inference (Rajpurkar et al., 2016)	92.64	accuracy
QQP	Quora Question Pairs	91.32	accuracy
RTE	Recognizing Textual Entailment (Dagan et al., 2006)	70.04	accuracy
SST-2	The Stanford Sentiment Treebank (Socher et al., 2013)	94.61	accuracy
STS-B	Semantic Textual Similarity Benchmark (Cer et al., 2017)	90.08	Pearson correlation

Table 2: Baseline results obtained for different GLUE tasks with RoBERTa-`base` and the relevant metric.

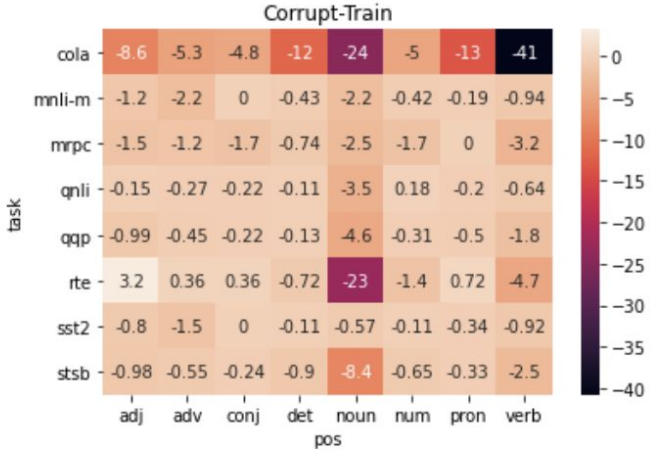

Figure 1: Impact of data corruption in the CORRUPT-TRAIN setting. The columns correspond to the removed word class and the rows to the GLUE tasks.

Results

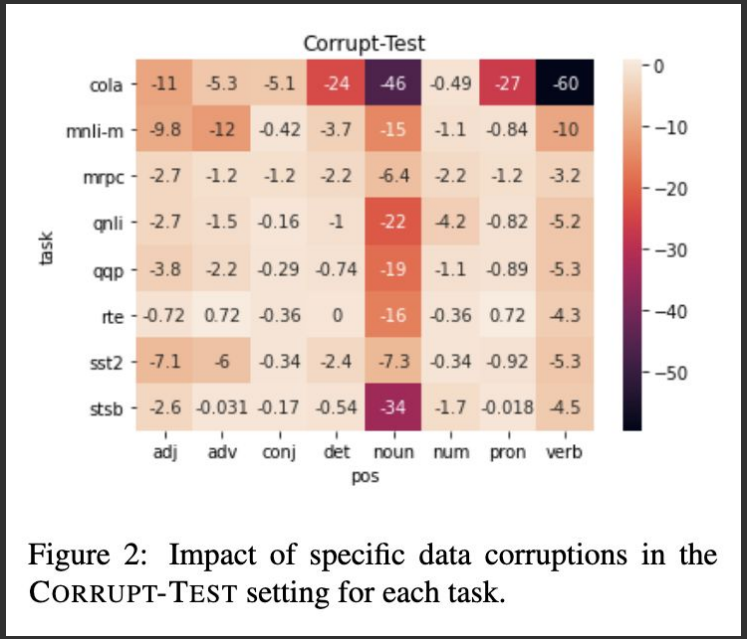

Figure 2: Impact of specific data corruptions in the CORRUPT-TEST setting for each task.

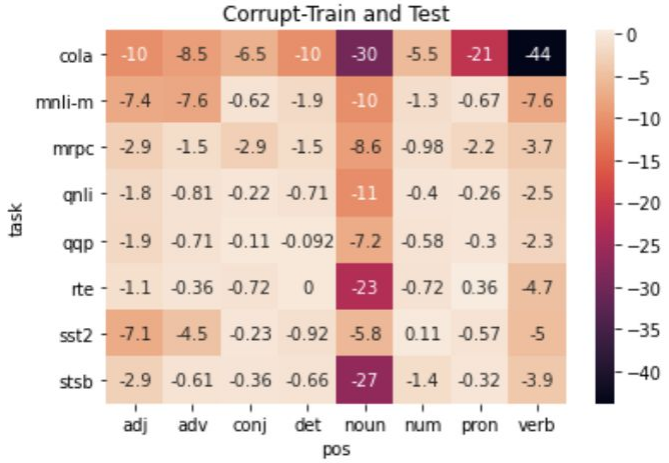

Figure 3: Impact of specific data corruptions in the CORRUPT-TRAIN AND TEST setting for each task.

Findings

The results confirm our earlier results for natural language inference tasks and show that those results hold also for the other GLUE tasks

- Our results indicate that understanding the meaning of utterances is not required for high performance in most GLUE tasks.
- This finding suggests caution in interpreting leaderboard results and drawing conclusions regarding the language understanding capabilities of the models.

How to fix language understanding testing?

Suggestions for future research

- There are various diagnostics that have been proposed in research.
- These are a good starting point when developing new NLU datasets and bechmark tasks:
  - Hypothesis only baseline for NLI datasets (Gururangan et al., 2018; Poliak et al., 2018)
  - Word-order shuffling (Pham et al., 2020)
  - Swapping sentences (e.g. in NLI swapping the premises and hypotheses) (Wang et al., 2019b)
  - Word class dropping (our proposed diagnostics)

Thank you!

Data and Code:

- https://github.com/Helsinki-NLP/nli-data-sanity-check
- https://github.com/Helsinki-NLP/nlu-dataset-diagnostics

Contact:

- Twitter: @aarnetalman
- Email: aarne.talman@helsinki.fi