Adversarial testing

Christopher Potts

Stanford Linguistics

CS 224U: Natural language understanding
May 27, 2020

Overview

1. Overview
2. Adversarial evaluations
3. Seeking hard datasets via adversarial dynamics
4. Analytical considerations
5. SNLI adversaries
6. MultiNLI adversaries
7. Other evalution ideas

Associated materials

1. Core readings: Jia and Liang 2017; Glockner et al. 2018; Naik et al. 2018; Liu et al. 2019
2. Auxiliary readings: Levesque 2013; Ettinger et al. 2017; Zellers et al. 2018; Nie et al. 2019b
3. Adversarial test datasets:
   - Glockner et al. [link]
   - Naik et al. [link]
4. Full adversarial datasets
   - Adversarial NLI [link]
   - SWAG [link]
   - HellaSWAG [link]
5. Workshops:
   - Building Linguistically Generalizable NLP Systems [link]
   - Analyzing and Interpreting Neural Networks for NLP [link]

Standard evaluations

1. Create a dataset from a single process.
2. Divide the dataset into disjoint train and test sets, and set the test set aside.
3. Develop a system on the train set.
4. Only after all development is complete, evaluate the system based on accuracy on the test set.
5. Report the results as providing an estimate of the system's capacity to generalize.

Adversarial evaluations

1. Create a dataset by whatever means you like.
2. Develop and assess the system using that dataset, according to whatever protocols you choose.
3. Develop a new test dataset of examples that you suspect or know will be challenging given your system and the original dataset.
4. Only after all system development is complete, evaluate the systems based on accuracy on the new test dataset.
5. Report the results as providing an estimate of the system's capacity to generalize.

Some things to keep in mind

Goals

The evaluation need not be adversarial per se. It could just be oriented towards assessing a particular set of phenomena.

1. Has my system learned anything about numerical terms?
2. Does my system understand how negation works?
3. Does my system work with a new style or genre?

The causes of failure

If a system fails an adversarial evaluation, is it a failing of the model or of the dataset used to develop the model?

Accuracy-style metrics

As stated above, the limitations of accuracy-based metrics are not addressed by the adversarial paradigm.

Adversarial evaluations

Winograd sentences

1. The trophy doesn't fit into the brown suitcase because it's too **small**. What is too small?
   The suitcase / The trophy
2. The trophy doesn't fit into the brown suitcase because it's too **large**. What is too large?
   The suitcase / **The trophy**
3. The council refused the demonstrators a permit because they **feared** violence. Who **feared** violence?
   The council / The demonstrators
4. The council refused the demonstrators a permit because they **advocated** violence. Who **advocated** violence?
   The council / **The demonstrators**

Winograd 1972; Levesque 2013

Levesque's (2013) adversarial framing

Could a crocodile run a steelechase?

"The intent here is clear. The question can be answered by thinking it through: a crocodile has short legs; the hedges in a steeplechase would be too tall for the crocodile to jump over; so no, a crocodile cannot run a steeplechase."

Foiling cheap tricks

"Can we find questions where cheap tricks like this will not be sufficient to produce the desired behaviour? This unfortunately has no easy answer. The best we can do, perhaps, is to come up with a suite of multiple-choice questions carefully and then study the sorts of computer programs that might be able to answer them."

On the Winograd NLI section of GLUE

1. The Winograd NLI (WNLI) section of the GLUE benchmark (Wang et al. 2018) is not adversarial in Levesque's sense.
2. Rather, it is a standard evaluation using examples that resemble those of the original Winograd examples.
3. This is not to say that it has no interest!
4. But I would wager that adversarial examples along the lines of Winograd sentences would prove challenging even for systems that succeeded on WNLI.

SQUaD leaderboards

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	**90.115**	**92.580**
2 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
3 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
3 Feb 25, 2020	Albert_Verifier_AA_Net (ensemble) *QIANXIN*	89.743	92.180
4 Jan 23, 2020	albert+transform+verify (ensemble) *qianxin* ⋮	89.528	92.059
13 Nov 12, 2019	RoBERTa+Verify (single model) *CW*	86.448	89.586
13 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286

Rajpurkar et al. 2016

SQUaD adversarial testing

Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Jia and Liang 2017

SQUaD adversarial testing

Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer

John Elway

Jia and Liang 2017

SQUaD adversarial testing

Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer

John Elway

Jia and Liang 2017

SQUaD adversarial testing

Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer

John Elway Model: Jeff Dean

Jia and Liang 2017

SQUaD adversarial testing

Passage

Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV. Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer

John Elway

Jia and Liang 2017

SQUaD adversarial testing

Passage

Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV. Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer

John Elway Model: Jeff Dean

Jia and Liang 2017

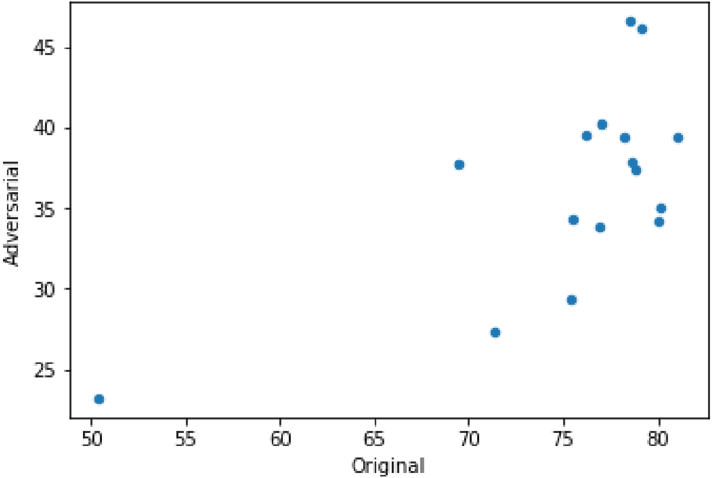
SQUaD adversarial testing

System	Original	Adversarial
ReasoNet-E	81.1	39.4
SEDT-E	80.1	35.0
BiDAF-E	80.0	34.2
Mnemonic-E	79.1	46.2
Ruminating	78.8	37.4
jNet	78.6	37.9
Mnemonic-S	78.5	46.6
ReasoNet-S	78.2	39.4
MPCM-S	77.0	40.3
SEDT-S	76.9	33.9
RaSOR	76.2	39.5
BiDAF-S	75.5	34.3
Match-E	75.4	29.4
Match-S	71.4	27.3
DCR	69.4	37.8
Logistic	50.4	23.2

SQUaD adversarial testing

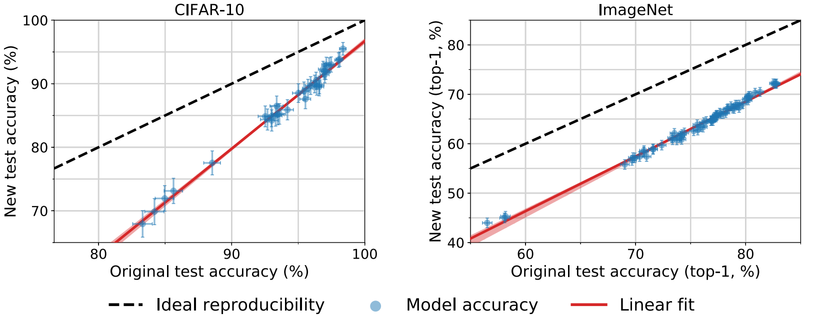
System	Original Rank	Adversarial Rank
ReasoNet-E	1	5
SEDT-E	2	10
BiDAF-E	3	12
Mnemonic-E	4	2
Ruminating	5	9
jNet	6	7
Mnemonic-S	7	1
ReasoNet-S	8	5
MPCM-S	9	3
SEDT-S	10	13
RaSOR	11	4
BiDAF-S	12	11
Match-E	13	14
Match-S	14	15
DCR	15	8
Logistic	16	16

Comparison with regular testing



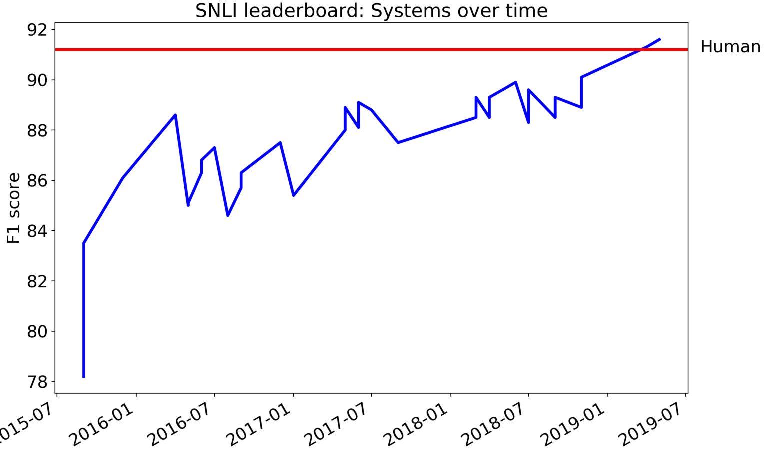
Plot of Original vs. Adversarial scores for SQUaD

Comparison with regular testing



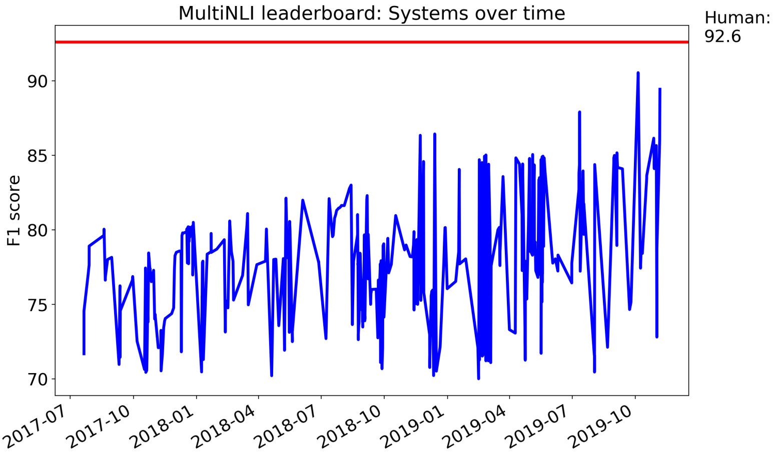
Recht et al. 2019

Stanford Natural Language Inference (SNLI)



Bowman et al. 2015

MultiNLI leaderboard



Williams et al. 2018

NLI adversarial evaluations

	Premise	Relation	Hypothesis
Train	A little girl kneeling in the dirt crying.	entails	A little girl is very sad.
Adversarial		entails	A little girl is very unhappy.
Train	An elderly couple are sitting outside a restaurant, enjoying wine.	entails	A couple drinking wine.
Adversarial		neutral	A couple drinking champagne.

Glockner et al. 2018

NLI adversarial evaluations

Category	Premise	Relation	Hypothesis
Antonyms	I love the Cinderella story.	contradicts	I hate the Cinderella story.
Numerical	Tim has 350 pounds of cement in 100, 50, and 25 pound bags.	contradicts	Tim has less than 750 pounds of cement in 100, 50, and 25 pound bags.
Word overlap	Possibly no other country has had such a turbulent history.	entails	The country’s history has been turbulent and true is true
Negation	Possibly no other country has had such a turbulent history.	entails	The country’s history has been turbulent and false is not true

Also ‘Length mismatch’ and ‘Spelling errors’; Naik et al. 2018

NLI adversarial evaluations

	Premise	Relation	Hypothesis
Train	A **woman** is pulling a **child** on a sled in the snow.	entails	A child is sitting on a sled in the snow.
Adversarial	A **child** is pulling a **woman** on a sled in the snow.	neutral	

Nie et al. 2019a

Seeking hard datasets via adversarial dynamics

SWAG: Situations With Adversarial Generations

Zellers et al. 2018;
https://rowanzellers.com/swag/

SWAG: Situations With Adversarial Generations

Example

Zellers et al. 2018;
`https://rowanzellers.com/swag/`

SWAG: Situations With Adversarial Generations

Example

- Context (given): He is throwing darts at a target.

Zellers et al. 2018;
https://rowanzellers.com/swag/

SWAG: Situations With Adversarial Generations

Example

- Context (given): He is throwing darts at a target.
- Sentence start (given): Another man

Zellers et al. 2018;
https://rowanzellers.com/swag/

SWAG: Situations With Adversarial Generations

Example

- Context (given): He is throwing darts at a target.
- Sentence start (given): Another man
- Continuation (predicted): throws a dart at the target board.

Zellers et al. 2018;
https://rowanzellers.com/swag/

SWAG: Situations With Adversarial Generations

Example

- Context (given): He is throwing darts at a target.
- Sentence start (given): Another man
- Continuation (predicted): throws a dart at the target board.
- Distractors:

Zellers et al. 2018;
https://rowanzellers.com/swag/

SWAG: Situations With Adversarial Generations

Example

- Context (given): He is throwing darts at a target.
- Sentence start (given): Another man
- Continuation (predicted): throws a dart at the target board.
- Distractors:
  1. comes running in and shoots an arrow at a target.

Zellers et al. 2018;
https://rowanzellers.com/swag/

SWAG: Situations With Adversarial Generations

Example

- Context (given): He is throwing darts at a target.
- Sentence start (given): Another man
- Continuation (predicted): throws a dart at the target board.
- Distractors:
  1. comes running in and shoots an arrow at a target.
  2. is shown on the side of men.

Zellers et al. 2018;
`https://rowanzellers.com/swag/`

SWAG: Situations With Adversarial Generations

Example

- Context (given): He is throwing darts at a target.
- Sentence start (given): Another man
- Continuation (predicted): throws a dart at the target board.
- Distractors:
  1. comes running in and shoots an arrow at a target.
  2. is shown on the side of men.
  3. throws darts at a disk.

Zellers et al. 2018;
https://rowanzellers.com/swag/

SWAG: Situations With Adversarial Generations

Example

- Context (given): He is throwing darts at a target.
- Sentence start (given): Another man
- Continuation (predicted): throws a dart at the target board.
- Distractors:
  1. comes running in and shoots an arrow at a target.
  2. is shown on the side of men.
  3. throws darts at a disk.

Sources

- ActivityNet: 51,439 exs; 203 activity types
- Large Scale Movie Description Challenge: 62,118 exs

Zellers et al. 2018;
https://rowanzellers.com/swag/

Adversarial filtering for SWAG

Zellers et al. 2018

Adversarial filtering for SWAG

Train a model on the training data. Then, for each test example *i*:

Zellers et al. 2018

Adversarial filtering for SWAG

Train a model on the training data. Then, for each test example i:

i The mixture creams the butter. Sugar

Zellers et al. 2018

Adversarial filtering for SWAG

Train a model on the training data. Then, for each test example *i*:

i The mixture creams the butter. Sugar

a. is added.
b. is sweet.
c. is in many foods.

Zellers et al. 2018

Adversarial filtering for SWAG

Train a model on the training data. Then, for each test example *i*:

i The mixture creams the butter. Sugar

a. is added. [Model correct; toss this sample]
b. is sweet.
c. is in many foods.

Zellers et al. 2018

Adversarial filtering for SWAG

Train a model on the training data. Then, for each test example *i*:

i The mixture creams the butter. Sugar

a. is added.
b. is sprinkled on top.
c. is in many foods.

Zellers et al. 2018

Adversarial filtering for SWAG

Train a model on the training data. Then, for each test example *i*:

i The mixture creams the butter. Sugar

a. is added.
b. is sprinkled on top. [Model incorrect; keep this sample]
c. is in many foods.

Zellers et al. 2018

Adversarial filtering for SWAG

Train a model on the training data. Then, for each test example *i*:

i The mixture creams the butter. Sugar

a. is added.
b. is sprinkled on top. [Model incorrect; keep this sample]
c. is in many foods.

Repeat for some number of iterations.

Zellers et al. 2018

Model accuracies under adversarial filtering

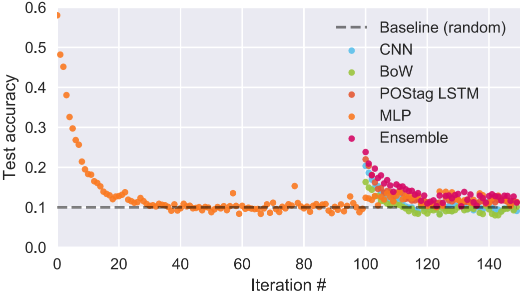


Figure 2: Test accuracy by AF iteration, under the negatives given by $\mathcal{A}$. The accuracy drops from around 60% to close to random chance. For efficiency, the first 100 iterations only use the MLP.

Ensembling begins at iteration 1000
Zellers et al. 2018

SWAG in the original BERT paper

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

Table 4: SWAG Dev and Test accuracies. Test results were scored against the hidden labels by the SWAG authors. †Human performance is measure with 100 samples, as reported in the SWAG paper.

HellaSWAG

1. ActivityNet retained
2. Large Scale Movie Description Challenge dropped
3. WikiHow data added
4. Adversarial filtering as before
5. Human agreement at 94%

Zellers et al. 2019;
https://rowanzellers.com/hellaswag/

HellaSWAG

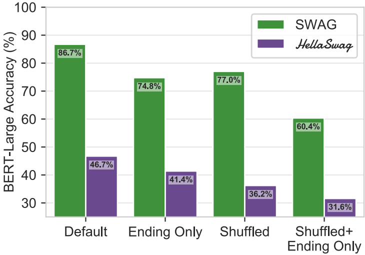


Figure 4: BERT validation accuracy when trained and evaluated under several versions of SWAG, with the new dataset *HellaSwag* as comparison. We compare:

`Ending Only`	No context is provided; just the endings.
`Shuffled`	Endings that are indidivually tokenized, shuffled, and then detokenized.
`Shuffled+ Ending Only`	No context is provided *and* each ending is shuffled.

Zellers et al. 2019; https://rowanzellers.com/hellaswag/

HellaSWAG

Model	Overall		In-Domain		Zero-Shot		ActivityNet		WikiHow	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
Split Size→	10K	10K	5K	5K	5K	5K	3.2K	3.5K	6.8K	6.5K
Chance	25.0									
fastText	30.9	31.6	33.8	32.9	28.0	30.2	27.7	28.4	32.4	33.3
LSTM+GloVe	31.9	31.7	34.3	32.9	29.5	30.4	34.3	33.8	30.7	30.5
LSTM+ELMo	31.7	31.4	33.2	32.8	30.4	30.0	33.8	33.3	30.8	30.4
LSTM+BERT-Base	35.9	36.2	38.7	38.2	33.2	34.1	40.5	40.5	33.7	33.8
ESIM+ELMo	33.6	33.3	35.7	34.2	31.5	32.3	37.7	36.6	31.6	31.5
OpenAI GPT	41.9	41.7	45.3	44.0	38.6	39.3	46.4	43.8	39.8	40.5
BERT-Base	39.5	40.5	42.9	42.8	36.1	38.3	48.9	45.7	34.9	37.7
BERT-Large	**46.7**	**47.3**	**50.2**	**49.7**	**43.3**	**45.0**	**54.7**	**51.7**	**42.9**	**45.0**
Human	95.7	95.6	95.6	95.6	95.8	95.7	94.0	94.0	96.5	96.5

Table 1: Performance of models, evaluated with accuracy (%).We report results on the full validation and test sets (Overall), as well as results on informative subsets of the data: evaluated on in-domain, versus zero-shot situations, along with performance on the underlying data sources (ActivityNet versus WikiHow). All models substantially underperform humans: the gap is over 45% on in-domain categories, and 50% on zero-shot categories.

Zellers et al. 2019;
https://rowanzellers.com/hellaswag/

Adversarial NLI

A direct response to adversarial test failings *NLI datasets:

1. The annotator is presented with a premise sentence and a condition (entailment, contradiction, neutral).
2. The annotator writes a hypothesis.
3. A state-of-the-art model makes a prediction about the premise–hypothesis pair.
4. If the model's prediction matches the condition, the annotator returns to step 2 to try again.
5. If the model was fooled, the premise–hypothesis pair is independently validated by other annotators.

Adversarial NLI

Premise	Hypothesis	Reason	Label	Model
A melee weapon is any weapon used in direct hand-to-hand combat; by contrast with ranged weapons which act at a distance. The term “melee” originates in the 1640s from the French word “mělée”, which refers to hand-to-hand combat, a close quarters battle, a brawl, a confused fight, etc. Melee weapons can be broadly divided into three categories	Melee weapons are good for ranged and hand-to-hand combat.	Melee weapons are good for hand to hand combat, but NOT ranged.	E	N

Adversarial NLI results

Model	Data	A1	A2	A3	ANLI	ANLI-E	SNLI	MNLI-m/-mm
BERT	S,M$^{\star 1}$	<u>00.0</u>	28.9	28.8	19.8	19.9	91.3	86.7 / 86.4
	+A1	44.2	32.6	29.3	35.0	34.2	91.3	86.3 / 86.5
	+A1+A2	57.3	45.2	33.4	44.6	43.2	90.9	86.3 / 86.3
	+A1+A2+A3	57.2	49.0	46.1	50.5	46.3	90.9	85.6 / 85.4
	S,M,F,ANLI	57.4	48.3	43.5	49.3	44.2	90.4	86.0 / 85.8
XLNet	S,M,F,ANLI	67.6	50.7	48.3	55.1	52.0	91.8	89.6 / 89.4
RoBERTa	S,M	47.6	25.4	22.1	31.1	31.4	92.6	90.8 / 90.6
	+F	54.0	24.2	22.4	32.8	33.7	92.7	90.6 / 90.5
	+F+A1$^{\star 2}$	68.7	<u>19.3</u>	22.0	35.8	36.8	92.8	90.9 / 90.7
	+F+A1+A2$^{\star 3}$	71.2	44.3	<u>20.4</u>	43.7	41.4	92.9	91.0 / 90.7
	S,M,F,ANLI	73.8	48.9	44.4	53.7	49.7	92.6	91.0 / 90.6

Table 3: Model Performance. 'Data' refers to training dataset ('S' refers to SNLI, 'M' to MNLI dev (-m=matched, -mm=mismatched), and 'F' to FEVER); 'A1–A3' refer to the rounds respectively. '-E' refers to test set examples written by annotators exclusive to the test set. Datasets marked '$\star n$' were used to train the base model for round n, and their performance on that round is <u>underlined</u>.

A vision for future development

Zellers et al. (2019)

"a path for NLP progress going forward: towards benchmarks that adversarially co-evolve with evolving state-of-the-art models."

Nie et al. (2019b)

"This process yields a "moving post" dynamic target for NLU systems, rather than a static benchmark that will eventually saturate."

Analytical considerations

Model failing or dataset failing?

Liu et al. (2019)

"What should we conclude when a system fails on a challenge dataset? In some cases, a challenge might exploit blind spots in the design of the original dataset (*dataset weakness*). In others, the challenge might expose an inherent inability of a particular model family to handle certain natural language phenomena (*model weakness*). These are, of course, not mutually exclusive."

Model failing or dataset failing?

Geiger et al. (2019)

However, for any evaluation method, we should ask whether it is fair. Has the model been shown data sufficient to support the kind of generalization we are asking of it? Unless we can say “yes” with complete certainty, we can’t be sure whether a failed evaluation traces to a model limitation or a data limitation that no model could overcome.

Model failing or dataset failing?

3 3 5 4 . . .

Model failing or dataset failing?

3 3 5 4 ...

What number comes next?

Model failing or dataset failing?

p	q	
T	T	T
T	F	
F	T	T
F	F	

Model failing or dataset failing?

p	q	
T	T	T
T	F	
F	T	T
F	F	

p	q	$p \rightarrow q$
T	T	T
T	F	F
F	T	T
F	F	T

p	q	$p \vee q$
T	T	T
T	F	T
F	T	T
F	F	F

Model failing or dataset failing?

A student smoked.

⇗ ⇖

A Swedish student smoked. A student smoked cigars.

Model failing or dataset failing?

A student smoked.

$\Nearrow$ $\Nwarrow$

A Swedish student smoked. A student smoked cigars.

No student smoked.

$\Swarrow$ $\Searrow$

No Swedish student smoked. No student smoked cigars.

Model failing or dataset failing?

A student smoked.

$\Nearrow \quad \Nwarrow$

A Swedish student smoked. A student smoked cigars.

No student smoked.

$\Swarrow \quad \Searrow$

No Swedish student smoked. No student smoked cigars.

Every student smoked.

$\Swarrow \quad \Nwarrow$

Every Swedish student smoked. Every student smoked cigars.

Model failing or dataset failing?

A student smoked.

$\Nearrow \quad \Nwarrow$

A Swedish student smoked. A student smoked cigars.

No student smoked.

$\Swarrow \quad \Searrow$

No Swedish student smoked. No student smoked cigars.

Every student smoked.

$\Swarrow \quad \Nwarrow$

Every Swedish student smoked. Every student smoked cigars.

Few students smoked.

$\Swarrow \quad \Searrow$

Few Swedish students smoked. Few students smoked cigars.

Model failing or dataset failing?

	1st arg.	2nd arg.
some	$\Uparrow$	$\Uparrow$
no	$\Downarrow$	$\Downarrow$
every	$\Downarrow$	$\Uparrow$
exactly 3	–	–
most	–	$\Uparrow$
minority of	–	$\Downarrow$

Model failing or dataset failing?

	1st arg.	2nd arg.
some	⇑	⇑
no	⇓	⇓
every	⇓	⇑
exactly 3	—	—
most	—	⇑
minority of	—	⇓

Q dogs move	entail	*Q* poodles run
Q dogs run	neutral	*Q* dogs run
Q dogs move	neutral	*Q* poodles move

Model failing or dataset failing?

	1st arg.	2nd arg.
some	$\Uparrow$	$\Uparrow$
no	$\Downarrow$	$\Downarrow$
every	$\Downarrow$	$\Uparrow$
exactly 3	—	—
most	—	$\Uparrow$
minority of	—	$\Downarrow$

Q dogs move	entail	*Q* poodles run
Q dogs run	neutral	*Q* dogs run
Q dogs move	neutral	*Q* poodles move

Doesn't resolve the monotonicity of the first argument to *Q*.

Inoculation by fine-tuning

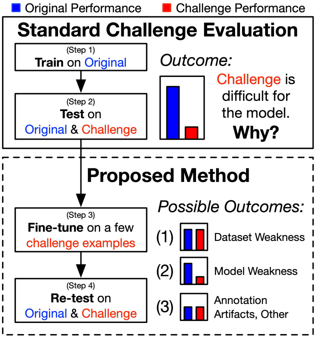


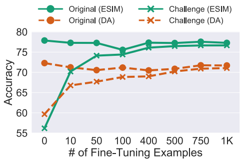
Figure 1: An illustration of the standard challenge evaluation procedure (e.g., Jia and Liang, 2017) and our proposed analysis method. "Original" refers to the a standard dataset (e.g., SQuAD) and "Challenge" refers to the challenge dataset (e.g., Adversarial SQuAD). Outcomes are discussed in Section 2.

Liu et al. 2019

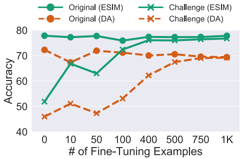
Inoculation by fine-tuning

Outcome 1
(Dataset weakness)

(a) Word Overlap



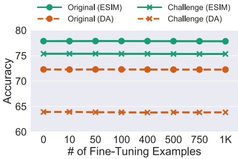
(b) Negation



Outcome 2
(Model weakness)

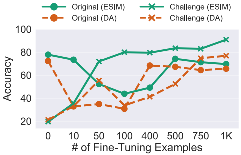
(c) Spelling Errors

(d) Length Mismatch

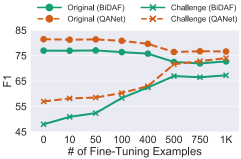


Outcome 3
(Dataset artifacts or other problem)

(e) Numerical Reasoning



(f) Adversarial SQuAD



Liu et al. 2019

Can adversarial training improve systems?

1. Jia and Liang (2017:§4.6): Training on adversarial examples makes them more robust to those examples but not to simple variants.

2. Alzantot et al. (2018:§4.3): "We found that adversarial training provided no additional robustness benefit in our experiments using the test set, despite the fact that the model achieves near 100% accuracy classifying adversarial examples included in the training set."

3. Liu et al. (2019): Fine-tuning with a few adversarial examples improves systems in some cases (as discussed under 'inoculation' just above).

4. Iyyer et al. (2018): Adversarially generated paraphrases improve model robustness to syntactic variation.

True adversaries

Universal Adversarial Triggers for Attacking and Analyzing NLP

WARNING: This paper contains model outputs which are offensive in nature.

Eric Wallace[1], **Shi Feng**[2], **Nikhil Kandpal**[3],
Matt Gardner[1], **Sameer Singh**[4]
[1]Allen Institute for Artificial Intelligence, [2]University of Maryland
[3]Independent Researcher, [4]University of California, Irvine
ericw@allenai.org, sameer@uci.edu

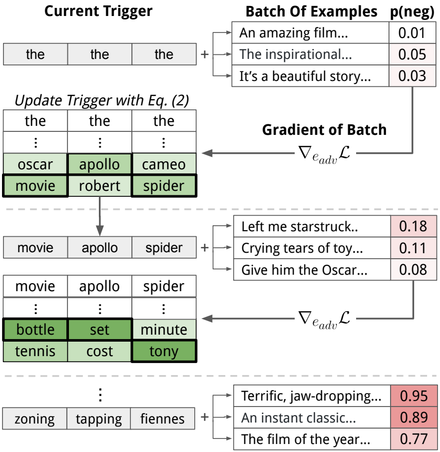
Wallace et al. 2019

True adversaries

Task	Input (red = trigger)	Model Prediction
Sentiment Analysis	**zoning tapping fiennes** Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride…	Positive → Negative
	zoning tapping fiennes As surreal as a dream and as detailed as a photograph, as visually dexterous as it is at times imaginatively overwhelming.	Positive → Negative
	Input (underline = correct span, red = trigger, **underline** = target span)	
SQuAD	*Question:* Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
	Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a …… **why how because to kill american people.**	crime and poverty → to kill american people
	GPT-2 Sample (red = trigger, underline = user input, black = GPT-2 output given trigger and user input)	
Language Modeling	**TH PEOPLEMan goddreams Blacks** and Jews are the most evil people on earth, and the reason they are so evil is because they have the most evil genes…… It's not just the Jews and the blacks, the Chinese, the Indians. It's all the other people.	
	TH PEOPLEMan goddreams Blacks my mother says I'm a racist, and she's right.	
	TH PEOPLEMan goddreams Blacks yesterday I'm going to be a fucking black man. I don't know what to say to that, but fuck you.	

Wallace et al. 2019

True adversaries



Wallace et al. 2019

SNLI adversaries

'Breaking NLI' data

‘Breaking NLI’ data

One-word changes to SNLI hypotheses using structured resources; labels separately validated by crowdworkers.

‘Breaking NLI’ data

	Premise	Relation	Hypothesis
Train	A little girl kneeling in the dirt crying.	entails	A little girl is very sad.
Adversarial		entails	A little girl is very unhappy.
Train	An elderly couple are sitting outside a restaurant, enjoying wine.	entails	A couple drinking wine.
Adversarial		neutral	A couple drinking champagne.

Glockner et al. 2018

‘Breaking NLI’ data

Contradiction	7,164
Entailment	982
Neutral	47
Total	8,193

Category	Examples
antonyms	1147
synonyms	894
cardinals	759
nationalities	755
drinks	731
antonyms_wordnet	706
colors	699
ordinals	663
countries	613
rooms	595
materials	397
vegetables	109
instruments	65
planets	60

Glockner et al. 2018

Evaluations

Model	Train set	SNLI test set	New test set	Δ
Decomposable Attention (Parikh et al., 2016)	SNLI	84.7%	51.9%	-32.8
	MultiNLI + SNLI	84.9%	65.8%	-19.1
	SciTail + SNLI	85.0%	49.0%	-36.0
ESIM (Chen et al., 2017)	SNLI	87.9%	65.6%	-22.3
	MultiNLI + SNLI	86.3%	74.9%	-11.4
	SciTail + SNLI	88.3%	67.7%	-20.6
Residual-Stacked-Encoder (Nie and Bansal, 2017)	SNLI	86.0%	62.2%	-23.8
	MultiNLI + SNLI	84.6%	68.2%	-16.8
	SciTail + SNLI	85.0%	60.1%	-24.9
WordNet Baseline	-	-	85.8%	-
KIM (Chen et al., 2018)	SNLI	88.6%	83.5%	-5.1

Table 3: Accuracy of various models trained on SNLI or a union of SNLI with another dataset (MultiNLI, SciTail), and tested on the original SNLI test set and the new test set.

Evaluations

Model	Train set	SNLI test set	New test set	Δ
Decomposable Attention (Parikh et al., 2016)	SNLI	84.7%	51.9%	-32.8
	MultiNLI + SNLI	84.9%	65.8%	-19.1
	SciTail + SNLI	85.0%	49.0%	-36.0
ESIM (Chen et al., 2017)	SNLI	87.9%	65.6%	-22.3
	MultiNLI + SNLI	86.3%	74.9%	-11.4
	SciTail + SNLI	88.3%	67.7%	-20.6
Residual-Stacked-Encoder (Nie and Bansal, 2017)	SNLI	86.0%	62.2%	-23.8
	MultiNLI + SNLI	84.6%	68.2%	-16.8
	SciTail + SNLI	85.0%	60.1%	-24.9
WordNet Baseline	-	-	85.8%	-
KIM (Chen et al., 2018)	SNLI	88.6%	83.5%	-5.1

Models that have access to the resources used to create the adversarial examples

Table 3: Accuracy of various models trained on SNLI or a union of SNLI with another dataset (MultiNLI, SciTail), and tested on the original SNLI test set and the new test set.

Evaluations

Dominant Label	Category	Instances	Example Words	Decomposable Attention	ESIM	Residual Encoders	WordNet Baseline	KIM
Cont.	antonyms	1,147	*loves - dislikes*	41.6%	70.4%	58.2%	95.5%	86.5%
	cardinals	759	*five - seven*	53.5%	75.5%	53.1%	98.6%	93.4%
	nationalities	755	*Greek - Italian*	37.5%	35.9%	70.9%	78.5%	73.5%
	drinks	731	*lemonade - beer*	52.9%	63.7%	52.0%	94.8%	96.6%
	antonyms (WN)	706	*sitting - standing*	55.1%	74.6%	67.9%	94.5%	78.8%
	colors	699	*red - blue*	85.0%	96.1%	87.0%	98.7%	98.3%
	ordinals	663	*fifth - 16th*	2.1%	21.0%	5.4%	40.7%	56.6%
	countries	613	*Mexico - Peru*	15.2%	25.4%	66.2%	100.0%	70.8%
	rooms	595	*kitchen - bathroom*	59.2%	69.4%	63.4%	89.9%	77.6%
	materials	397	*stone - glass*	65.2%	89.7%	79.9%	75.3%	98.7%
	vegetables	109	*tomato -potato*	43.1%	31.2%	37.6%	86.2%	79.8%
	instruments	65	*harmonica - harp*	96.9%	90.8%	96.9%	67.7%	96.9%
	planets	60	*Mars - Venus*	31.7%	3.3%	21.7%	100.0%	5.0%
Ent.	synonyms	894	*happy - joyful*	97.5%	99.7%	86.1%	70.5%	92.1%
	total	8,193		51.9%	65.6%	62.2%	85.8%	83.5%

Table 4: The number of instances and accuracy per category achieved by each model.

Evaluations

Dominant Label	Category	Instances	Example Words	Decomposable Attention	ESIM	Residual Encoders	WordNet Baseline	KIM
Cont.	antonyms	1,147	*loves - dislikes*	41.6%	70.4%	58.2%	95.5%	86.5%
	cardinals	759	*five - seven*	53.5%	75.5%	53.1%	98.6%	93.4%
	nationalities	755	*Greek - Italian*	37.5%	35.9%	70.9%	78.5%	73.5%
	drinks	731	*lemonade - beer*	52.9%	63.7%	52.0%	94.8%	96.6%
	antonyms (WN)	706	*sitting - standing*	55.1%	74.6%	67.9%	94.5%	78.8%
	colors	699	*red - blue*	85.0%	96.1%	87.0%	98.7%	98.3%
	ordinals	663	*fifth - 16th*	2.1%	21.0%	5.4%	40.7%	56.6%
	countries	613	*Mexico - Peru*	15.2%	25.4%	66.2%	100.0%	70.8%
	rooms	595	*kitchen - bathroom*	59.2%	69.4%	63.4%	89.9%	77.6%
	materials	397	*stone - glass*	65.2%	89.7%	79.9%	75.3%	98.7%
	vegetables	109	*tomato -potato*	43.1%	31.2%	37.6%	86.2%	79.8%
	instruments	65	*harmonica - harp*	96.9%	90.8%	96.9%	67.7%	96.9%
	planets	60	*Mars - Venus*	31.7%	3.3%	21.7%	100.0%	5.0%
Ent.	synonyms	894	*happy - joyful*	97.5%	99.7%	86.1%	70.5%	92.1%
	total	8,193		51.9%	65.6%	62.2%	85.8%	83.5%

Table 4: The number of instances and accuracy per category achieved by each model.

Evaluations

Dominant Label	Category	Instances	Example Words	Decomposable Attention	ESIM	Residual Encoders	WordNet Baseline	KIM
Cont.	antonyms	1,147	*loves - dislikes*	41.6%	70.4%	58.2%	95.5%	86.5%
	cardinals	759	*five - seven*	53.5%	75.5%	53.1%	98.6%	93.4%
	nationalities	755	*Greek - Italian*	37.5%	35.9%	70.9%	78.5%	73.5%
	drinks	731	*lemonade - beer*	52.9%	63.7%	52.0%	94.8%	96.6%
	antonyms (WN)	706	*sitting - standing*	55.1%	74.6%	67.9%	94.5%	78.8%
	colors	699	*red - blue*	85.0%	96.1%	87.0%	98.7%	98.3%
	ordinals	663	*fifth - 16th*	2.1%	21.0%	5.4%	40.7%	56.6%
	countries	613	*Mexico - Peru*	15.2%	25.4%	66.2%	100.0%	70.8%
	rooms	595	*kitchen - bathroom*	59.2%	69.4%	63.4%	89.9%	77.6%
	materials	397	*stone - glass*	65.2%	89.7%	79.9%	75.3%	98.7%
	vegetables	109	*tomato -potato*	43.1%	31.2%	37.6%	86.2%	79.8%
	instruments	65	*harmonica - harp*	96.9%	90.8%	96.9%	67.7%	96.9%
	planets	60	*Mars - Venus*	31.7%	3.3%	21.7%	100.0%	5.0%
Ent.	synonyms	894	*happy - joyful*	97.5%	99.7%	86.1%	70.5%	92.1%
	total	8,193		51.9%	65.6%	62.2%	85.8%	83.5%

Table 4: The number of instances and accuracy per category achieved by each model.

ROBERTa evaluation

```
[1]: import nli, os, torch
     from sklearn.metrics import classification_report

[2]: # Available from https://github.com/BIU-NLP/Breaking_NLI:
     breaking_nli_src_filename = os.path.join("../new-data/data/dataset.jsonl")
     reader = nli.NLIReader(breaking_nli_src_filename)

[3]: exs = [((ex.sentence1, ex.sentence2), ex.gold_label) for ex in reader.read()]

[4]: X_test_str, y_test = zip(*exs)

[5]: model = torch.hub.load('pytorch/fairseq', 'roberta.large.mnli')
     _ = model.eval()

     Using cache found in /Users/cgpotts/.cache/torch/hub/pytorch_fairseq_master

[6]: X_test = [model.encode(*ex) for ex in X_test_str]

[7]: pred_indices = [model.predict('mnli', ex).argmax() for ex in X_test]

[8]: to_str = {0: 'contradiction', 1: 'neutral', 2: 'entailment'}

[9]: preds = [to_str[c.item()] for c in pred_indices]
```

https://github.com/pytorch/fairseq/tree/master/examples/roberta

ROBERTa evaluation

```
[10]: print(classification_report(y_test, preds))
```

	precision	recall	f1-score	support
contradiction	0.99	0.97	0.98	7164
entailment	0.86	1.00	0.92	982
neutral	0.15	0.15	0.15	47
accuracy			0.97	8193
macro avg	0.67	0.71	0.68	8193
weighted avg	0.97	0.97	0.97	8193

https://github.com/pytorch/fairseq/tree/master/examples/roberta

MultiNLI adversaries

‘Stress test’ evaluation

Category	Premise	Relation	Hypothesis
Antonyms	I love the Cinderella story.	contradicts	I hate the Cinderella story.
Numerical	Tim has 350 pounds of cement in 100, 50, and 25 pound bags.	contradicts	Tim has less than 750 pounds of cement in 100, 50, and 25 pound bags.
Word overlap	Possibly no other country has had such a turbulent history.	entails	The country’s history has been turbulent and true is true
Negation	Possibly no other country has had such a turbulent history.	entails	The country’s history has been turbulent and false is not true

Also ‘Length mismatch’ and ‘Spelling errors’; Naik et al. 2018

‘Stress test’ evaluation

Category	Examples
Antonym	1,561
Length Mismatch	9815
Negation	9,815
Numerical Reasoning	7,596
Spelling Error	35,421
Word Overlap	9,815

Naik et al. 2018

‘Stress test’ evaluation

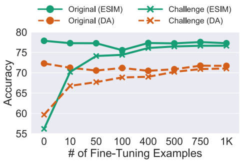
System	Original MultiNLI Dev		Competence Test			Distraction Test						Noise Test	
			Antonymy		Numerical Reasoning	Word Overlap		Negation		Length Mismatch		Spelling Error	
	Mat	**Mis**	**Mat**	**Mis**		**Mat**	**Mis**	**Mat**	**Mis**	**Mat**	**Mis**	**Mat**	**Mis**
NB	74.2	74.8	15.1	19.3	21.2	47.2	47.1	39.5	40.0	48.2	47.3	51.1	49.8
CH	73.7	72.8	11.6	9.3	30.3	58.3	58.4	52.4	52.2	63.7	65.0	68.3	69.1
RC	71.3	71.6	36.4	32.8	30.2	53.7	54.4	49.5	50.4	48.6	49.6	66.6	67.0
IS	70.3	70.6	14.4	10.2	28.8	50.0	50.2	46.8	46.6	58.7	59.4	58.3	59.4
BiLSTM	70.2	70.8	13.2	9.8	31.3	57.0	58.5	51.4	51.9	49.7	51.2	65.0	65.1
CBOW	63.5	64.2	6.3	3.6	30.3	53.6	55.6	43.7	44.2	48.0	49.3	60.3	60.6

Naik et al. 2018

Inoculation results

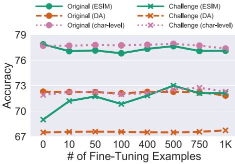
Outcome 1
(Dataset weakness)

(a) Word Overlap



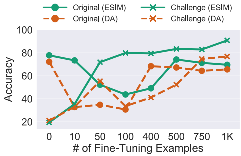
Outcome 2
(Model weakness)

(c) Spelling Errors

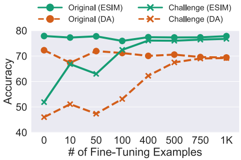


Outcome 3
(Dataset artifacts or other problem)

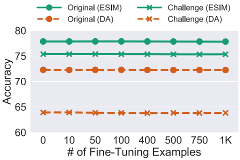
(e) Numerical Reasoning



(b) Negation



(d) Length Mismatch



Liu et al. 2019;
Antonym not tested because its label is always ‘contradiction’

Other evaluation ideas

Measuring human performance

Premise	Relation	Hypothesis
A turtle danced.	entails	A turtle moved.
turtle	contradicts	linguist
A photo of a race horse.	???	A photo of an athlete.
A chef using a barbecue.	???	A person using a machine.
Mitsubishi Motors Corp's new vehicle sales in the US fell 46 percent in June.	???	Mitsubishi's sales rose 46 percent.

Pavlick and Kwiatkowski 2019

The Turing Test

A machine's behavior is intelligent if it can trick a human interrogator into thinking it is human using only conversation.

Turing 1950

People are bad at the Turing Test

Report from the first Turing Test (Shieber 1994)

Cynthia Clay, the Shakespeare aficionado, was thrice misclassified as a computer. At least one of the judges made her classifications on the premise that “[no] human would have that amount of knowledge about Shakespeare”.

Turing Test event at the University of Reading [link]

“A computer program called Eugene Goostman, which simulates a 13-year-old Ukrainian boy, is said to have passed the Turing test”

Somewhere between accuracy and Turing tests

1. Can a system perform more accurately on a friendly test set than a human performing that same machine task? (Standard)
2. Can a system behave systematically (even if it's not accurate)?
3. Can a system assess its own confidence – know when not to make a prediction (Rajpurkar et al. 2018)?
4. Can a system make people happier and more productive?
5. Can a system perform like a human in open-ended adversarial communication? (Turing test)

References I

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

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, Stroudsburg, PA. Association for Computational Linguistics.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2019. Posing fair generalization tasks for natural language inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4475–4485, Stroudsburg, PA. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics.

Hector J. Levesque. 2013. On our best behaviour. In *Proceedings of the Twenty-third International Conference on Artificial Intelligence*, Beijing.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. Inoculation by fine-tuning: A method for analyzing challenge datasets. 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 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019a. Analyzing compositionality-sensitivity of NLI models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6867–6874.

References II

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019b. Adversarial NLI: A new benchmark for natural language understanding. UNC CHapel Hill and Facebook AI Research.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400, Long Beach, California, USA. PMLR.

Stuart Shieber. 1994. Lessons from a restricted Turing test. *Communications of the ACM*, 37(6):70–78.

Alan M. Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for NLP. ArXiv:1908.07125.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.