Analysis methods in NLP: Adversarial testing

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding

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 system based on accuracy on the new test dataset.
5. Report the results as providing an estimate of the system's capacity to generalize.

A bit of history

VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

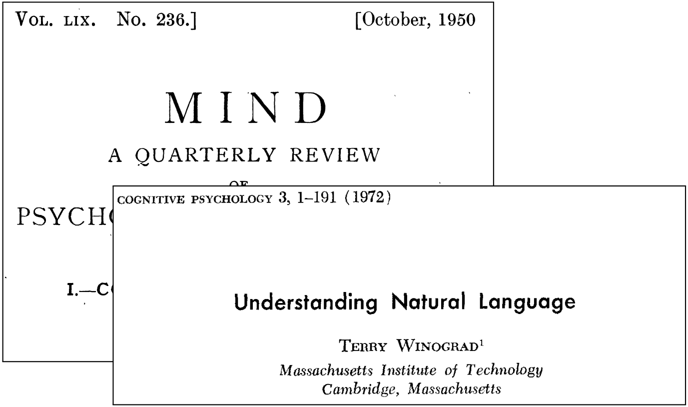
OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

A bit of history


Vol. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

PSYCH

I.—C

COGNITIVE PSYCHOLOGY 3, 1–191 (1972)

Understanding Natural Language

Terry Winograd[1]

Massachusetts Institute of Technology
Cambridge, Massachusetts

A bit of history

VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCH

I.—C

COGNITIVE PSYCHOLOGY 3, 1–191 (1972)

Understanding Natural Language

TERRY

Massachusetts I

Cambridg

On our best behaviour

Hector J. Levesque

Dept. of Computer Science

University of Toronto

Toronto, Ontario

Canada M5S 3A6

hector@cs.toronto.edu

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.”

Analytical considerations

What can adversarial testing tell us?

(And what can’t it tell us)?

No need to be too adversarial

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?

Metrics

The limitations of accuracy-based metrics are generally left unaddressed by the adversarial paradigm.

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?

Inoculation by fine-tuning

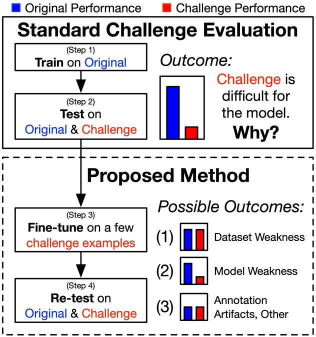


Figure 1: An illustration of the standard challenge evaluation procedure (e.g., Jia and Liang, 2017) and our proposed analysis method. "Original" refers to the a standard dataset (e.g., SQuAD) and "Challenge" refers to the challenge dataset (e.g., Adversarial SQuAD). Outcomes are discussed in Section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

Liu et al. 2019

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 Leland Stanford 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 Leland Stanford 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: Leland Stanford

Jia and Liang 2017

SQUaD adversarial testing

Passage

Quarterback Leland Stanford 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 Leland Stanford 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: Leland Stanford

Jia and Liang 2017

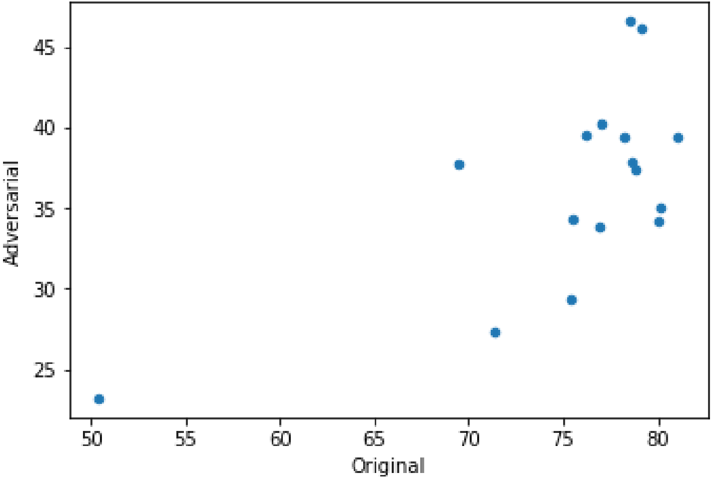
SQUaD adversarial testing

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

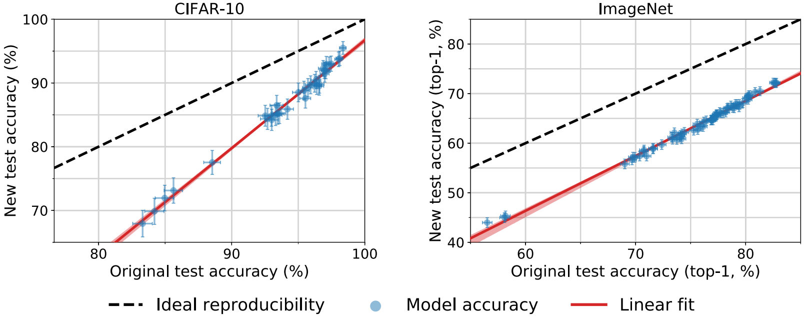
System	Original Rank	Adversarial Rank
ReasoNet-E	1	5
SEDT-E	2	10
BiDAF-E	3	12
Mnemonic-E	4	2
Ruminating	5	9
jNet	6	7
Mnemonic-S	7	1
ReasoNet-S	8	5
MPCM-S	9	3
SEDT-S	10	13
RaSOR	11	4
BiDAF-S	12	11
Match-E	13	14
Match-S	14	15
DCR	15	8
Logistic	16	16

Comparison with regular testing



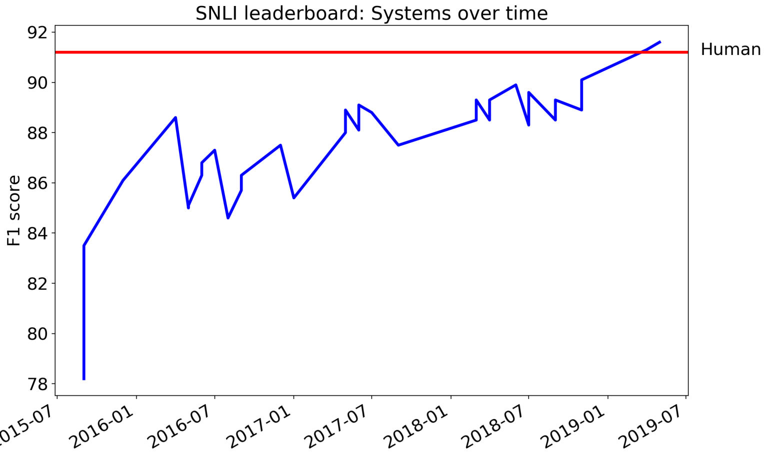
Plot of Original vs. Adversarial scores for SQUaD

Comparison with regular testing



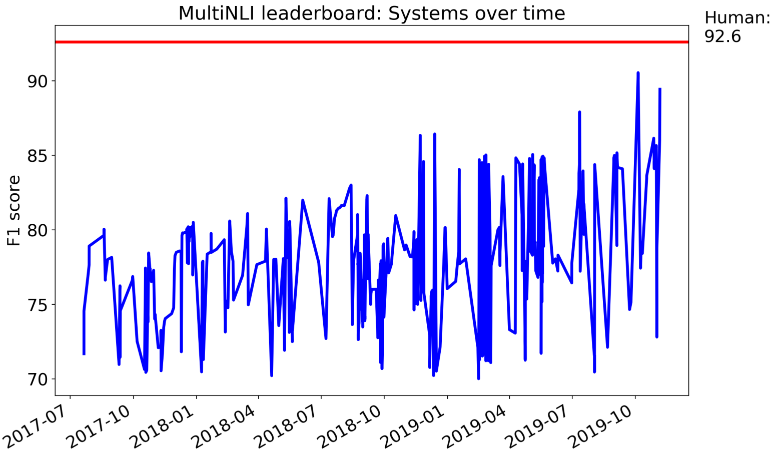
Recht et al. 2019

Example: NLI



Bowman et al. 2015

Example: NLI



Bowman et al. 2015

An SNLI adversarial evaluation

	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

An SNLI adversarial evaluation

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.

An SNLI adversarial evaluation

RoBERTA-MNLI, off-the-shelf

```
[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]
```

An SNLI adversarial evaluation

RoBERTA-MNLI, off-the-shelf

```
[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
```

A MultiNLI adversarial 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

A MultiNLI adversarial evaluation

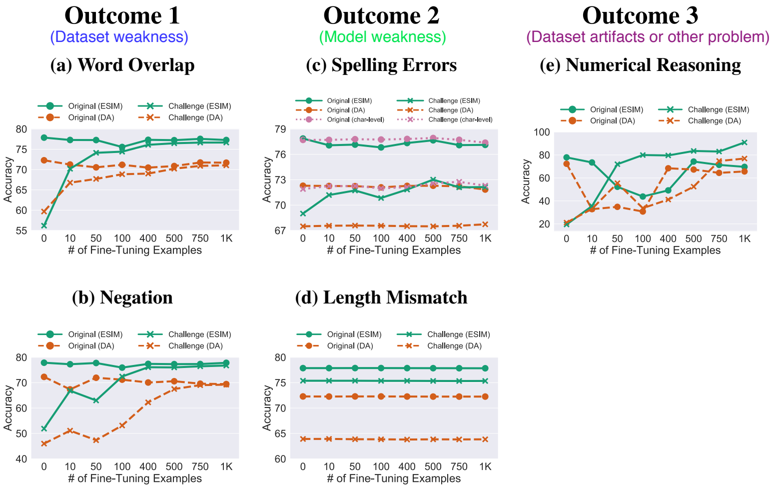
Category	Examples
Antonym	1,561
Length Mismatch	9815
Negation	9,815
Numerical Reasoning	7,596
Spelling Error	35,421
Word Overlap	9,815

Naik et al. 2018

A MultiNLI adversarial evaluation

System	Original MultiNLI Dev		Competence Test: Antonymy		Competence Test: Numerical Reasoning	Distraction Test: Word Overlap		Distraction Test: Negation		Distraction Test: Length Mismatch		Noise Test: 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

A MultiNLI adversarial evaluation



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

References I

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.

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.

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.

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.

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