Adversarial testing in natural language understanding

Christopher Potts

Stanford Linguistics

XCS224U: Natural language understanding
July 8, 2020

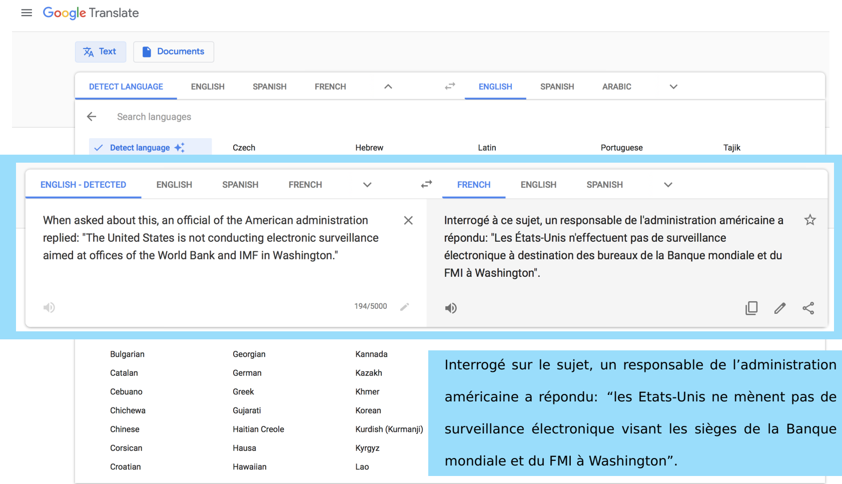
Overview

1. A golden age for Natural Language Understanding (NLU)
2. A peek behind the curtain
3. Adversarial testing
4. Coursework

A golden age for NLU

Artificial assistants

Translation



Interrogé sur le sujet, un responsable de l'administration américaine a répondu: "les Etats-Unis ne mènent pas de surveillance électronique visant les sièges de la Banque mondiale et du FMI à Washington".

Image captioning

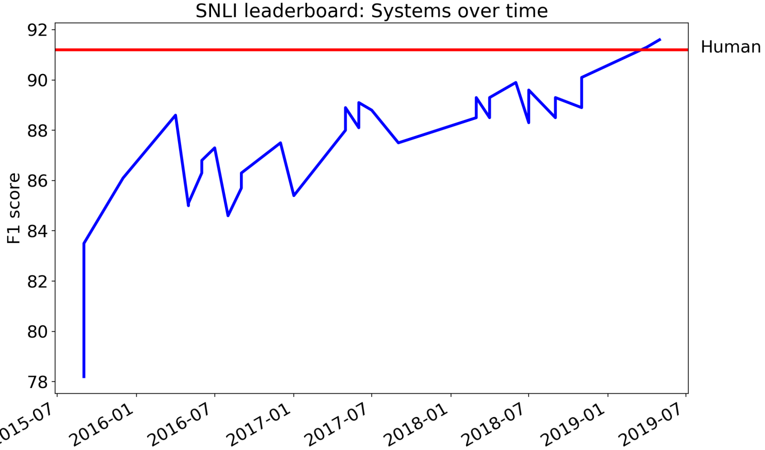
Sutskever et al. 2014

Watson wins Jeopardy (2011)

Natural Language Inference (NLI)

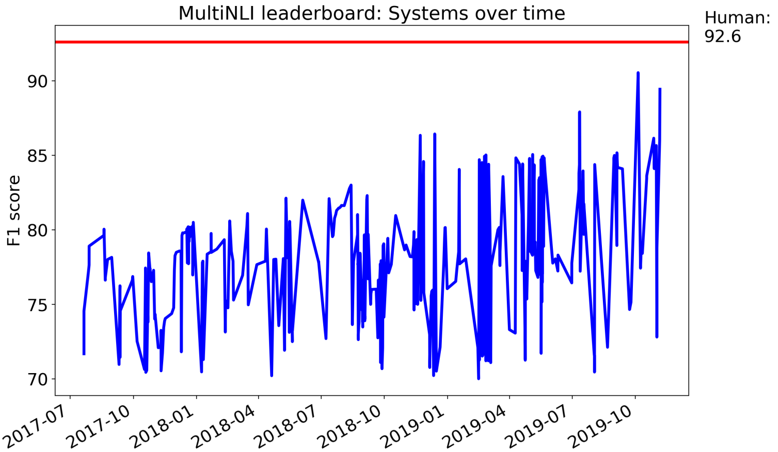
Premise	Relation	Hypothesis
A turtle danced.	entails	A turtle moved.
Every reptile danced.	neutral	A turtle ate.
Some turtles walk.	contradicts	No turtles move.

Stanford Natural Language Inference (SNLI)



Bowman et al. 2015

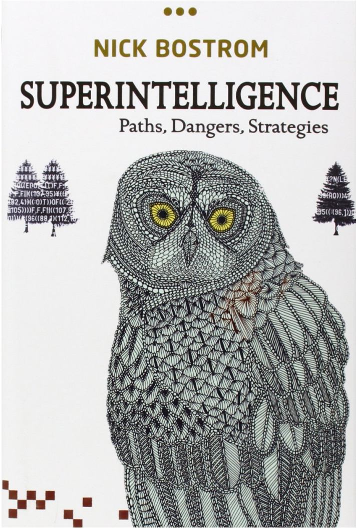
MultiNLI



Williams et al. 2018

“Superhuman” performance on other tasks

- NIST 2000 Switchboard Speech Recognition
- English-to-German WMT19 News Translation
- Stanford Question Answering Dataset (SQuAD)
- General Language Understanding Evaluation (GLUE)
- ...


NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies

A peek behind the curtain

The promise of artificial assistants

You: Any good burger joints around here?

Siri: I found a number of burger restaurants near you.

You: Hmm. How about tacos?

Apple: [Siri remembers that you asked about restaurants. so it will look for Mexican restaurants in the neighborhood. And Siri is proactive, so it will question you until it finds what you’re looking for.]

Slide idea from Marie de Marneffe

SIRI on The Colbert Show

Colbert: For the love of God, the cameras are on, give me something?

Siri: What kind of place are you looking for? Camera stores or churches?

[. . .]

Colbert: I don’t want to search for anything! I want to write the show!

Siri: Searching the Web for “search for anything. I want to write the shuffle.”

Slide idea from Marie de Marneffe

Translation: Garbage in, fluent text out?

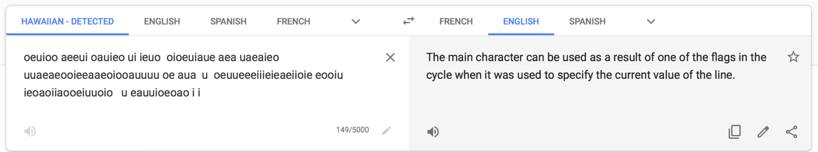


Image captioning

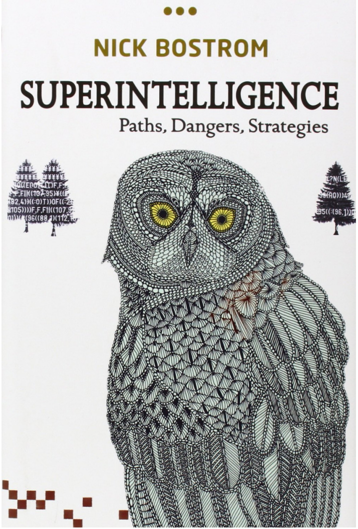
Sutskever et al. 2014

Watson gets confused

- Answer: Grasshoppers eat it.
- Watson: kosher

Class	Forbidden kinds
Mammals	Carnivores; animals that do not chew the cud (e.g., the pig); animals that do not have cloven hooves (e.g., the camel, the hare, the horse and the hyrax); bats
Birds	Birds of prey; scavengers
Reptiles and amphibians	All
Water animals	All non-fish. Among fish, all those that do not have both fins and scales
Insects	All, except particular types of locust or grasshopper that, according to most, cannot be identified today

Two perspectives



Adversarial testing

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

NLI adversarial testing

Premise	Relation	Hypothesis
A turtle danced.	entails	A turtle moved.
Every reptile danced.	neutral	A turtle ate.
Some turtles walk.	contradicts	No turtles move.

NLI adversarial testing

	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

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

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

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

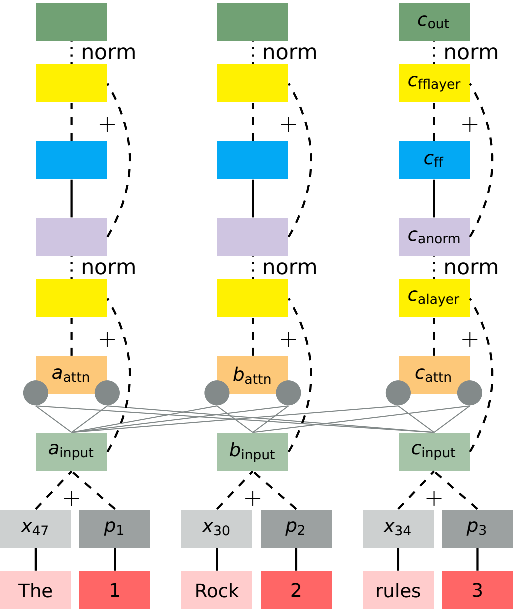
Glockner et al. 2018

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

Glockner et al. 2018

Transformer-based models BERT, ROBERTa, ELECTRA, XLNet, . . .



$$c_{\text{out}} = \frac{c_{\text{fflayer}} - \textbf{mean}(c_{\text{fflayer}})}{\textbf{std}(c_{\text{fflayer}}) + \epsilon}$$

$$c_{\text{fflayer}} = c_{\text{anorm}} + \textbf{Dropout}(c_{\text{ff}})$$

$$c_{\text{ff}} = \textbf{ReLU}(c_{\text{anorm}} W_1 + b_1) W_2 + b_2$$

$$c_{\text{anorm}} = \frac{c_{\text{alayer}} - \textbf{mean}(c_{\text{alayer}})}{\textbf{std}(c_{\text{alayer}}) + \epsilon}$$

$$c_{\text{alayer}} = \textbf{Dropout}\left(c_{\text{attn}} + c_{\text{input}}\right)$$

$$c_{\text{attn}} = \textbf{sum}\left([\alpha_1 a_{\text{input}}, \alpha_2 b_{\text{input}}]\right)$$

$$\alpha = \textbf{softmax}(\tilde{\alpha})$$

$$\tilde{\alpha} = \left[\frac{{c_{\text{input}}}^{\top} a_{\text{input}}}{\sqrt{d_k}}, \frac{{c_{\text{input}}}^{\top} b_{\text{input}}}{\sqrt{d_k}}\right]$$

$$c_{\text{input}} = x_{34} + p_3$$

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

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

The earlier adversaries didn't get above 0.75 accuracy!

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

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.

Nie et al. 2019

Adversarial NLI results

Model	Data	A1	A2	A3	ANLI	ANLI-E	SNLI	MNLI-m/-mm
BERT	S,M$^{\star 1}$	00.0	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	19.3	22.0	35.8	36.8	92.8	90.9 / 90.7
	+F+A1+A2$^{\star 3}$	71.2	44.3	20.4	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

Nie et al. 2019

Coursework

High-level summary

Topics

1. Vector-space models
2. Sentiment analysis
3. Relation extraction
4. NLI
5. Grounding
6. Contextual word representations
7. Adversarial testing
8. Methods and metrics

Assignments/bake-offs

1. Word similarity
2. Relation extraction with distant supervision
3. Word-level entailment
4. Generating color descriptions in context

Final projects

1. Literature review
2. Experiment protocol
3. Short video presentation
4. Final paper

Assignments and bake-offs

1. Each assignment culminates in a bake-off: an informal competition in which you enter your original model.
2. The assignments ask you to build baseline systems to inform your own model design, and to build your original model.
3. Winning bake-off entries earn extra credit.
4. Rationale for all this: exemplify best practices for NLU projects. (Let us know where we're not living up to this!)

Assign/Bake-off: Word-level entailment

Train		
turtle	animal	1
turtle	desk	0
ingredient	element	1
pain	joint	0
⋮		
Test		
dog	mammal	1
grenade	cycling	0
⋮		

Assign/Bake-off: Word-level entailment

Train		
turtle	animal	1
turtle	desk	0
ingredient	element	1
pain	joint	0
⋮		
Test		
dog	mammal	1
grenade	cycling	0
⋮		

Train and test have disjoint *vocabs*.

Assign/Bake-off: Word-level entailment

[1]:
```
import numpy as np
import torch.nn as nn
from torch_shallow_neural_classifier import TorchShallowNeuralClassifier
import utils
```

[2]:
```
def glove_vec(w):
    """Return `w`'s GloVe representation if available, else return
    a random vector."""
    return GLOVE.get(w, utils.randvec(w, n=50))
```

[3]:
```
def vec_concatenate(u, v):
    """Concatenate np.array instances `u` and `v` into a new np.array."""
    return np.concatenate((u, v))
```

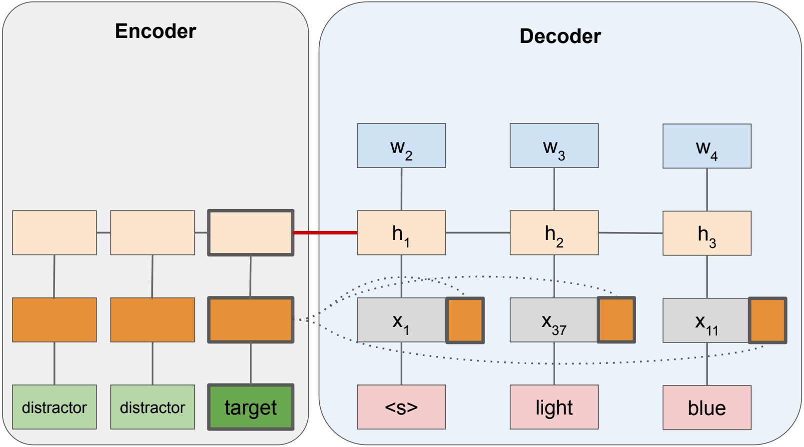
[4]:
```
class TorchDeepNeuralClassifier(TorchShallowNeuralClassifier):
    def __init__(self, dropout_prob=0.7, **base_kwargs):
        self.dropout_prob = dropout_prob
        super().__init__(**base_kwargs)

    def build_graph(self):
        """Adapt the following network to include an additional hidden
        layer with dropout regularization applied to it."""
        return nn.Sequential(
            nn.Linear(self.input_dim, self.hidden_dim),
            self.hidden_activation,
            nn.Linear(self.hidden_dim, self.n_classes_))
```

Assign/Bake-off: Contextual color describers

Context	Utterance
	blue
	The darker blue one
	dull pink not the super bright one
	Purple
	blue

Monroe et al. 2017, 2018

Assign/Bake-off: Contextual color describers



Monroe et al. 2017, 2018

Wrap-up

1. This is the most exciting moment ever in history for doing NLU!
2. This course will give you **hands-on** experience with a wide range of challenging NLU problems.
3. A mentor from the teaching team will guide you through the project assignments – there are many examples of these projects becoming important publications.
4. Central goal: to make you the best – most insightful and responsible – NLU researcher and practitioner wherever you go next.

Thanks!

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.

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.

Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Will Monroe, Jennifer Hu, Andrew Jong, and Christopher Potts. 2018. Generating bilingual pragmatic color references. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2155–2165, Stroudsburg, PA. Association for Computational Linguistics.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

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.