Natural Language Inference:

Challenges and Opportunities

Vered Shwartz

Natural Logic Meets Machine Learning, June 2021

Microsoft

Microsoft DeBERTa surpasses human performance on ...

... of NLU tasks, including question answering, natural language inference, co-reference resolution, word sense disambiguation, and others.

Jan 6, 2021

Analytics India Magazine

Facebook Inch Closer To General-Purpose Intelligence ...

The Facebook AI researchers started with the question: “Could a Transformer trained for natural language inference on textual input also ...

Mar 3, 2021

Microsoft

Microsoft DeBERTa surpasses human performance on ...

... of NLU tasks, including question answering, natural language inference, co-reference resolution, word sense disambiguation, and others.

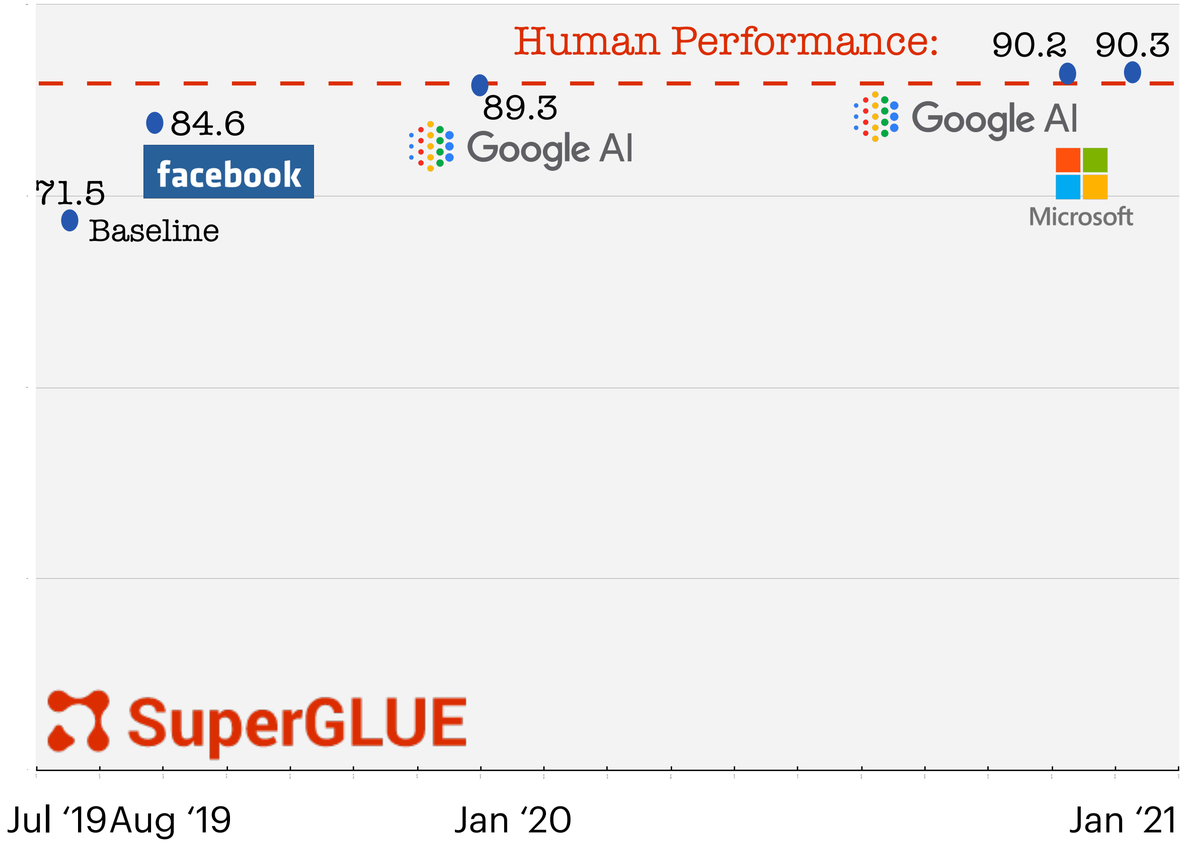
Jan 6, 2021

Analytics India Magazine

Facebook Inch Closer To General-Purpose Intelligence ...

The Facebook AI researchers started with the question: "Could a Transformer trained for natural language inference on textual input also ...

Mar 3, 2021



Microsoft

Microsoft DeBERTa surpasses human performance on ...

... of NLU tasks, including question answering, natural language inference, co-reference resolution, word sense disambiguation, and others.

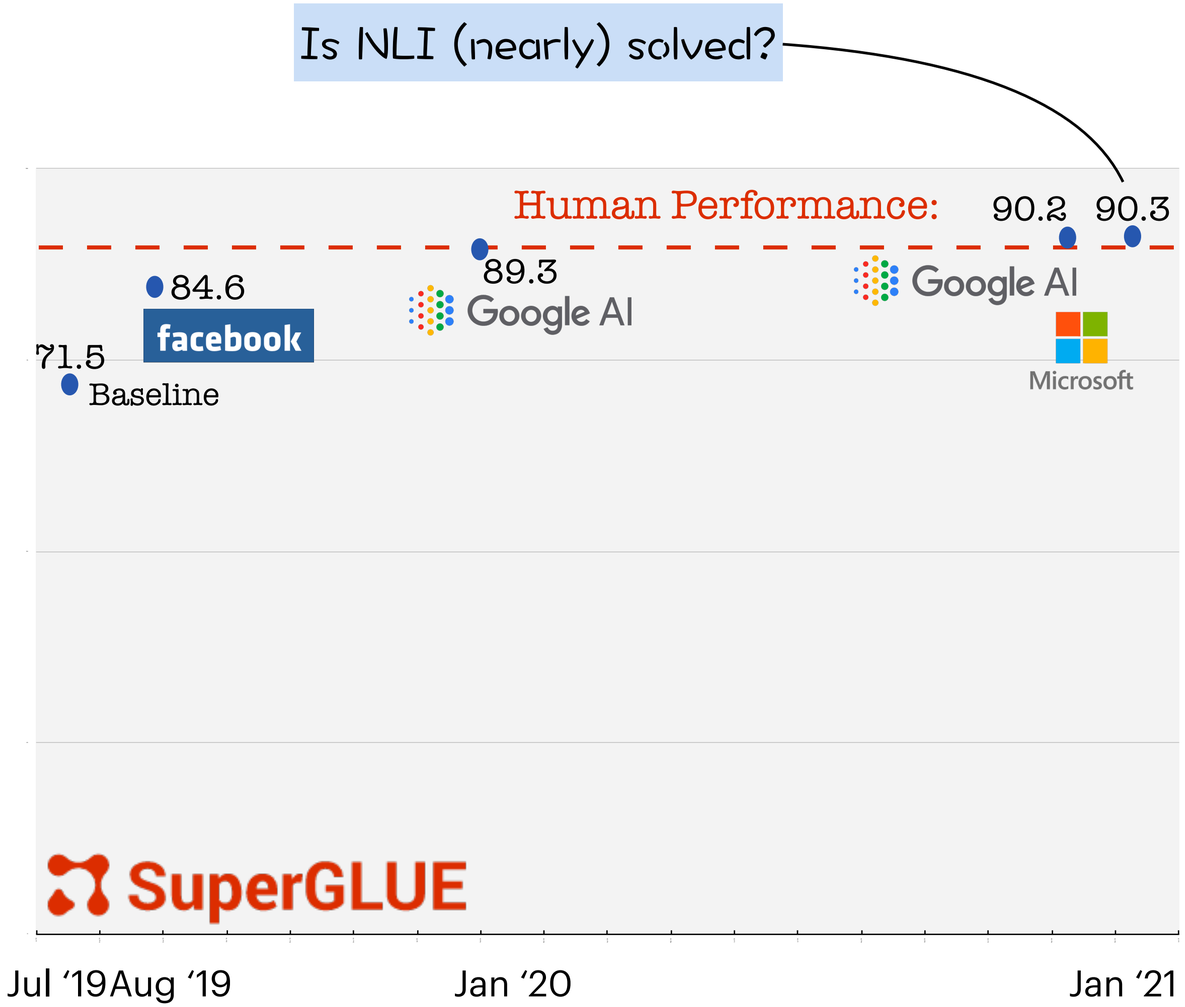
Jan 6, 2021

Analytics India Magazine

Facebook Inch Closer To General-Purpose Intelligence ...

The Facebook AI researchers started with the question: "Could a Transformer trained for natural language inference on textual input also ...

Mar 3, 2021

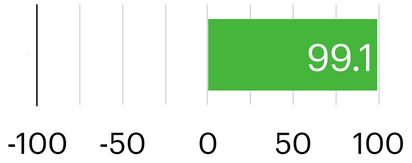


I baked a chocolate cake but accidentally used regular instead of self-rising flour.

https://demo.allennlp.org/textual-entailment

I baked a chocolate cake but accidentally used regular instead of self-rising flour.

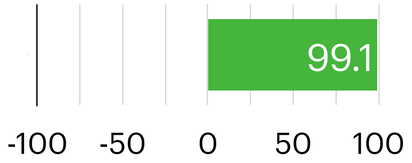
I've made a mistake.



https://demo.allennlp.org/textual-entailment

I baked a chocolate cake but accidentally used regular instead of self-rising flour.

I've made a mistake.



The cake turned out flat.

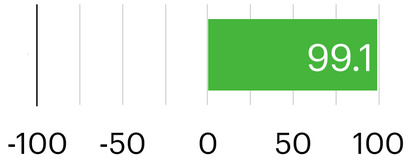
The cake was difficult to cut.

I baked a bad chocolate cake.

https://demo.allennlp.org/textual-entailment

I baked a chocolate cake but accidentally used regular instead of self-rising flour.

I've made a mistake.



The cake turned out flat.

The cake was difficult to cut.

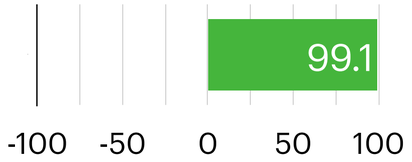
I baked a bad chocolate cake.

The cake was soft and tasty.

https://demo.allennlp.org/textual-entailment

I baked a chocolate cake but accidentally used regular instead of self-rising flour.

I've made a mistake.



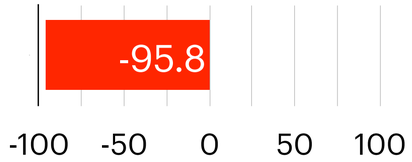
The cake turned out flat.

The cake was difficult to cut.

I baked a bad chocolate cake.

The cake was soft and tasty.

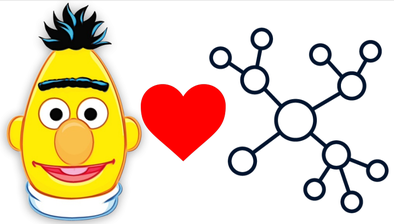
I've baked a chocolate rock.



https://demo.allennlp.org/textual-entailment

Outline

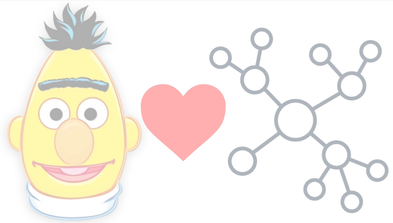
Limitations of neural NLI Models



Incorporating symbolic knowledge into neural NLI models

NLI is too easy? What's next?

Limitations of neural NLI Models

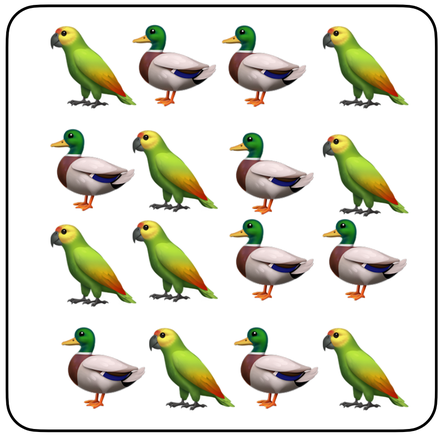


Incorporating symbolic knowledge into neural NLI models

NLI is too easy? What's next?

#1 Learning Paradigm

Dataset

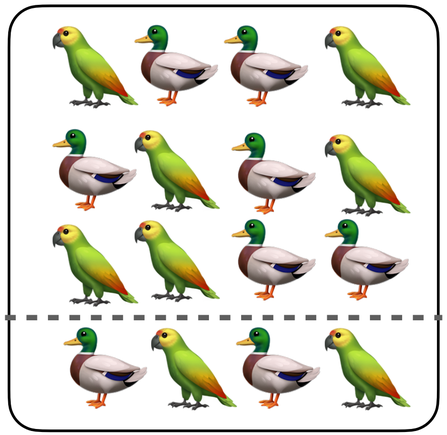


Labels

parrot, duck, duck, parrot...

#1 Learning Paradigm

Dataset

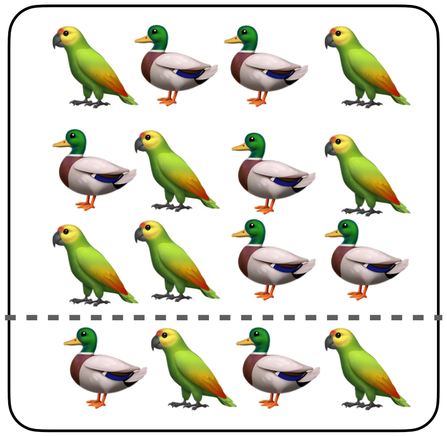


Labels

parrot, duck, duck, parrot...

#1 Learning Paradigm

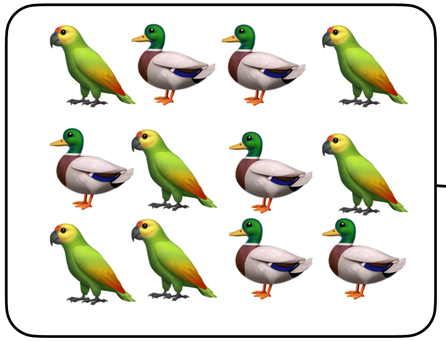
Dataset



Labels

parrot, duck, duck, parrot...

Training set



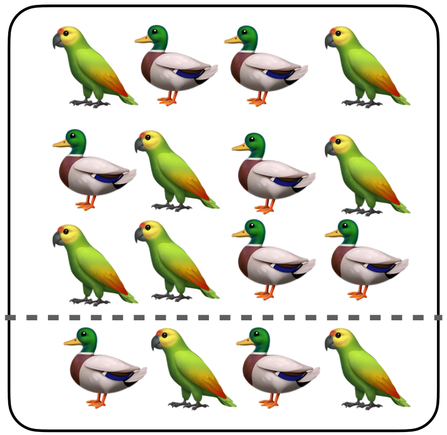
Labels

parrot, duck, duck, parrot...

Classifier

#1 Learning Paradigm

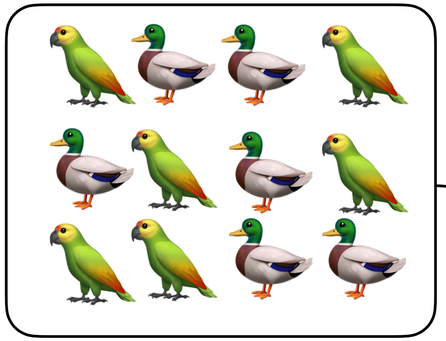
Dataset



Labels

parrot, duck, duck, parrot...

Training set

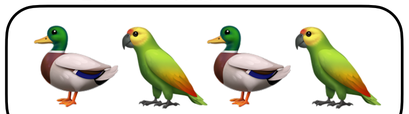


Labels

parrot, duck, duck, parrot...

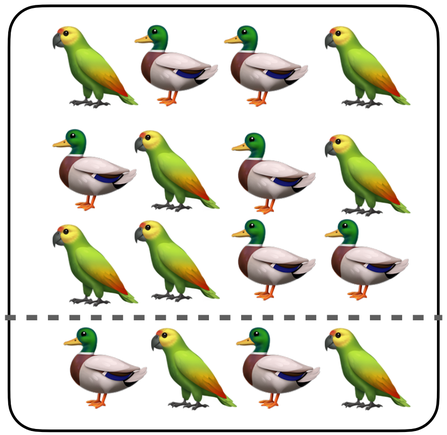
Classifier

Test set



#1 Learning Paradigm

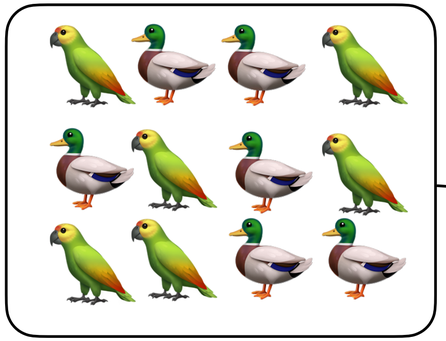
Dataset



Labels

parrot, duck, duck, parrot...

Training set



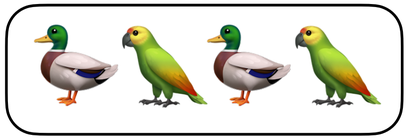
Labels

parrot, duck, duck, parrot...

Classifier

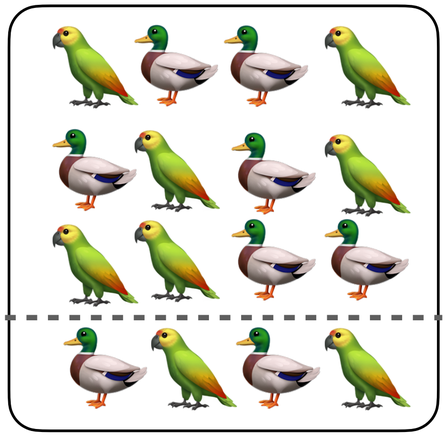
duck, parrot, duck, parrot...

Test set



#1 Learning Paradigm

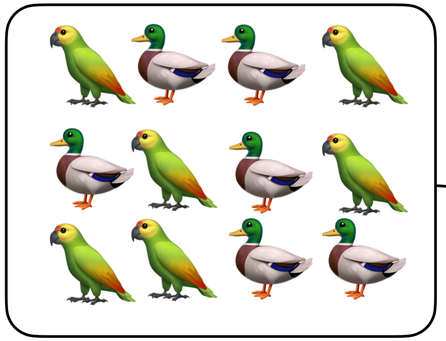
Dataset



Labels

parrot, duck, duck, parrot...

Training set



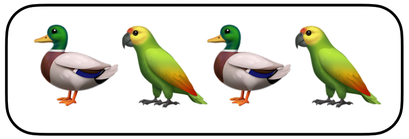
Labels

parrot, duck, duck, parrot...

Classifier

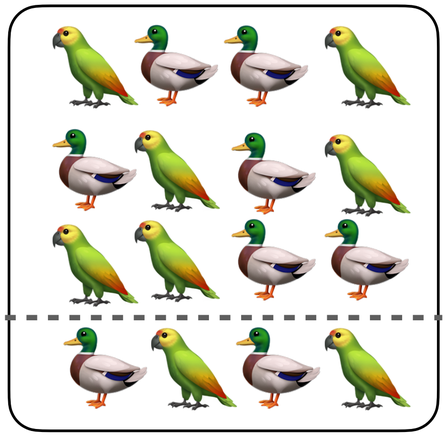
duck, parrot, duck, parrot...

Test set



#1 Learning Paradigm

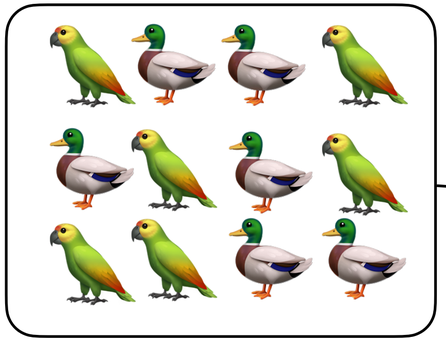
Dataset



Labels

parrot, duck, duck, parrot...

Training set



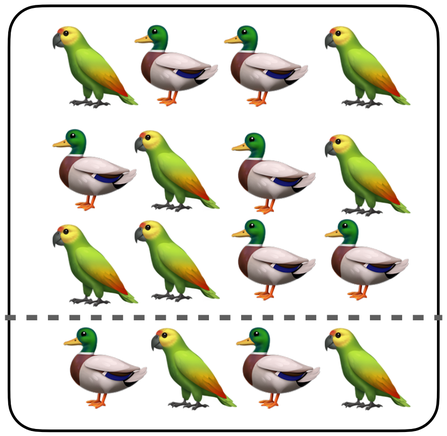
Labels

parrot, duck, duck, parrot...

Classifier

#1 Learning Paradigm

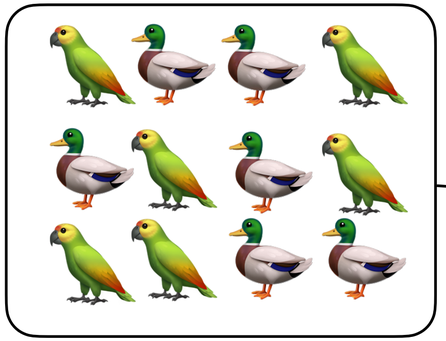
Dataset



Labels

parrot, duck, duck, parrot...

Training set



Labels

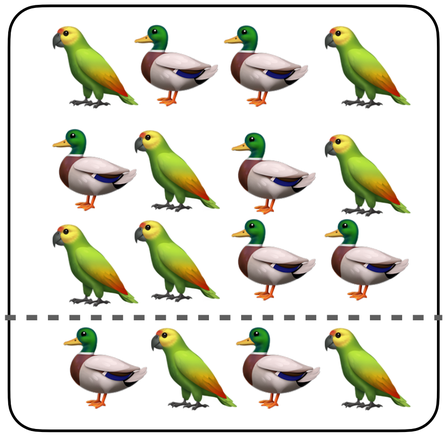
parrot, duck, duck, parrot...

Yellow head → parrot

Classifier

#1 Learning Paradigm

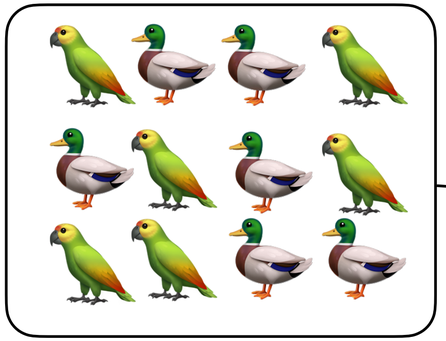
Dataset



Labels

parrot, duck, duck, parrot...

Training set

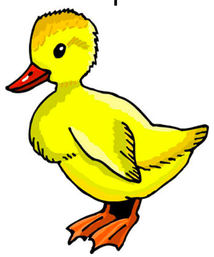


Labels

parrot, duck, duck, parrot...

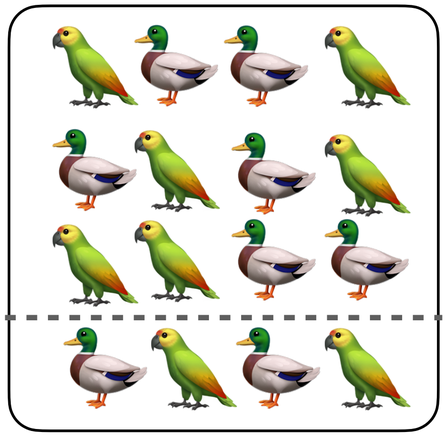
Yellow head → parrot

Classifier → parrot



#1 Learning Paradigm

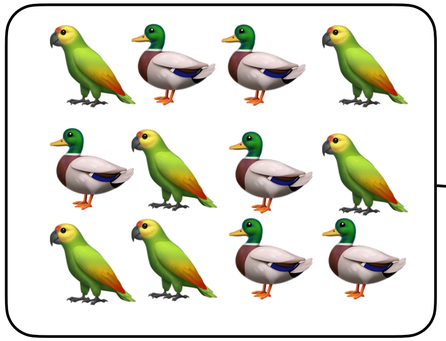
Dataset



Labels

parrot, duck, duck, parrot...

Training set



Labels

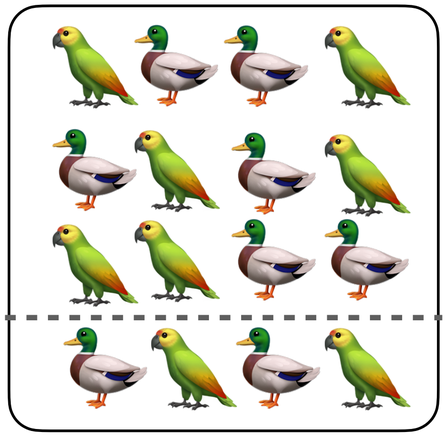
parrot, duck, duck, parrot...

Green body → parrot,
Gray body → duck

Classifier

#1 Learning Paradigm

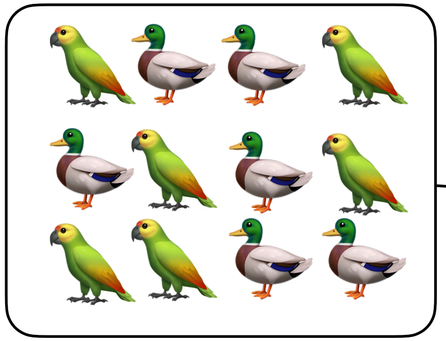
Dataset



Labels

parrot, duck, duck, parrot...

Training set



Labels

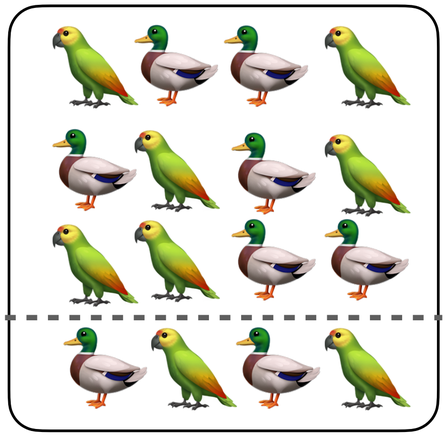
parrot, duck, duck, parrot...

Green body → parrot,
Gray body → duck

Classifier → duck

#1 Learning Paradigm

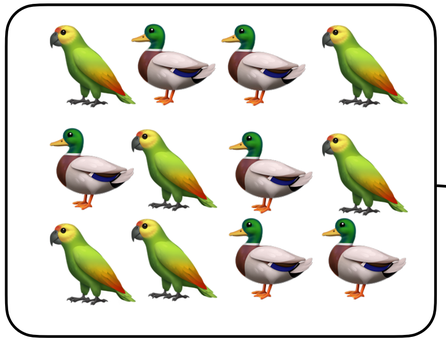
Dataset



Labels

parrot, duck, duck, parrot...

Training set



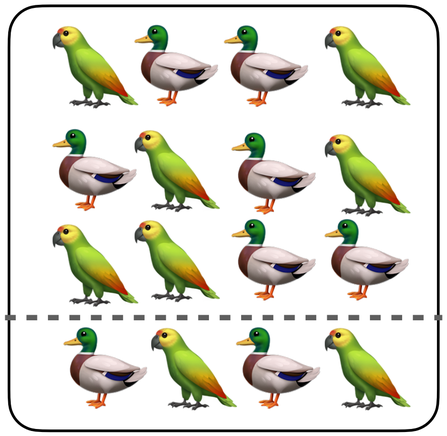
Labels

parrot, duck, duck, parrot...

Classifier

#1 Learning Paradigm

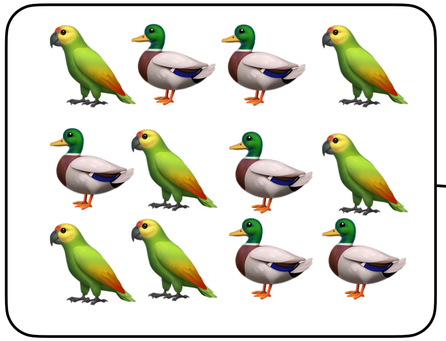
Dataset



Labels

parrot, duck, duck, parrot...

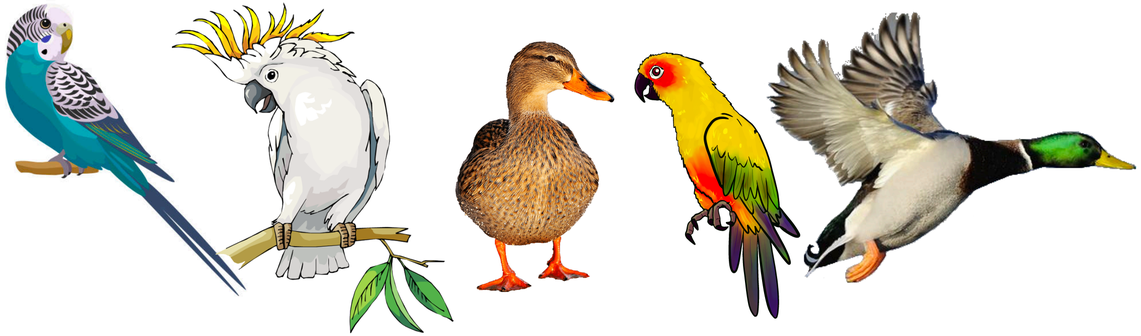
Training set



Labels

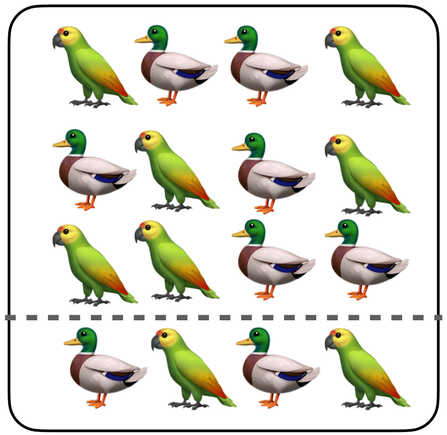
parrot, duck, duck, parrot...

Classifier



#1 Learning Paradigm

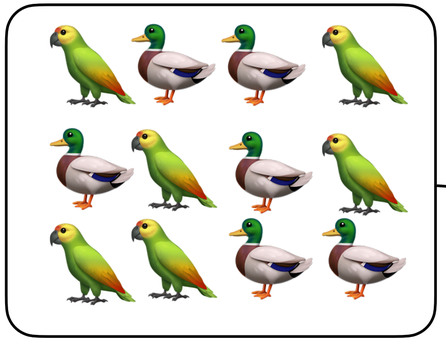
Dataset



Labels

parrot, duck, duck, parrot...

Training set

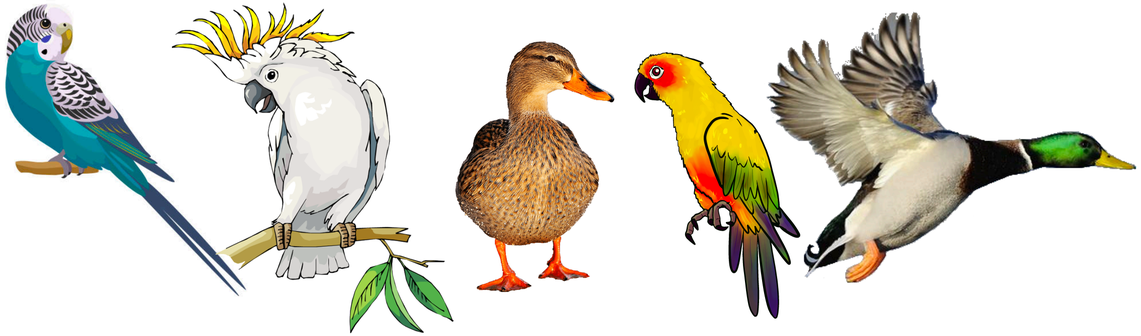


Labels

parrot, duck, duck, parrot...

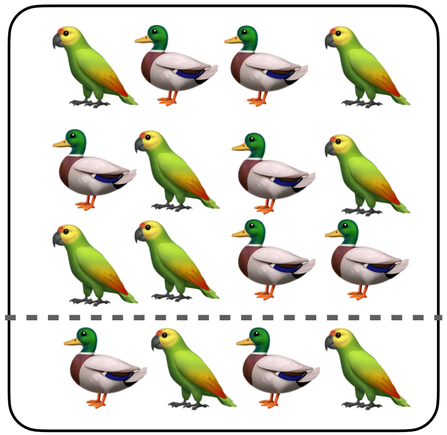
Classifier

1. Spurious correlations



#1 Learning Paradigm

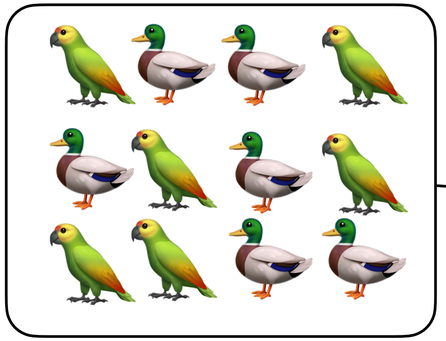
Dataset



Labels

parrot, duck, duck, parrot...

Training set

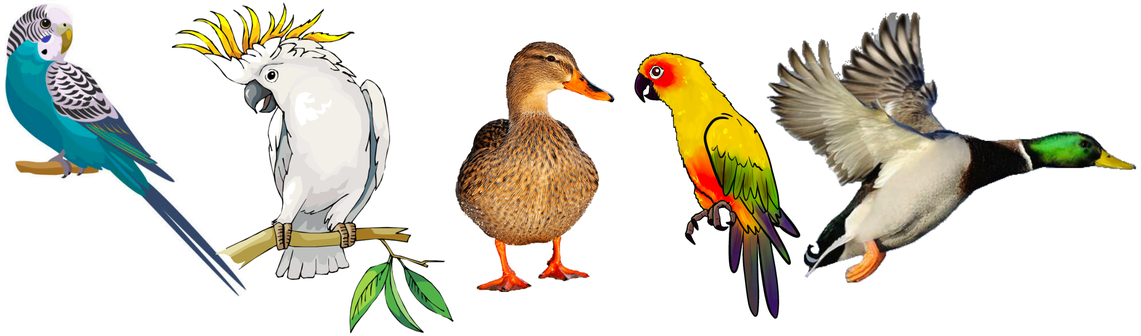


Labels

parrot, duck, duck, parrot...

Classifier

1. Spurious correlations
2. Train & test from the same distribution



#1 Learning Paradigm

Annotation artifacts, hypothesis-only baseline
(**Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018; Tsuchiya, 2018**)

#1 Learning Paradigm

Annotation artifacts, hypothesis-only baseline
(**Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018; Tsuchiya, 2018**)

p: I only had a soup but it was very filling.

h: I didn't eat a salad.

🤖: contradiction (98.2%)

#1 Learning Paradigm

Annotation artifacts, hypothesis-only baseline
(**Gururangan, Swayamdipta, et al., 2018; Poliak et al., 2018; Tsuchiya, 2018**)

p: I only had a soup but it was very filling.

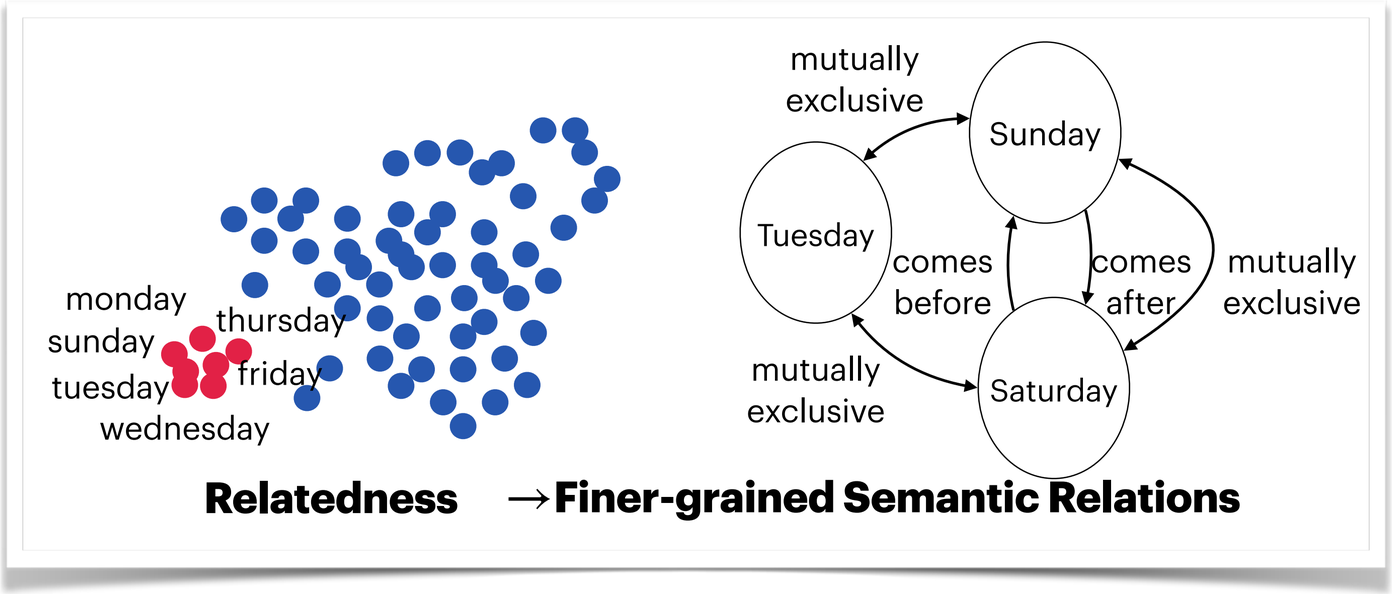
h: I didn't eat a salad.

🤖: contradiction (98.2%)

NLI models rely on syntactic heuristics
(**McCoy et al., 2019**)

#2 Representations

#2 Representations



#2 Representations

P: Charlie will visit his mother in London on **Wednesday** evening.

H: Charlie will visit his mother in London on **Thursday** evening.

#2 Representations

2018

(ELMo-based Decomposable Attention)

P: Charlie will visit his mother in London on **Wednesday** evening.

H: Charlie will visit his mother in London on **Thursday** evening.

https://demo.allennlp.org/textual-entailment

#2 Representations

2018

(ELMo-based Decomposable Attention)

P: Charlie will visit his mother in London on **Wednesday** evening.
H: Charlie will visit his mother in London on **Thursday** evening.

Prediction: Entailment (94.1%) ✘

https://demo.allennlp.org/textual-entailment

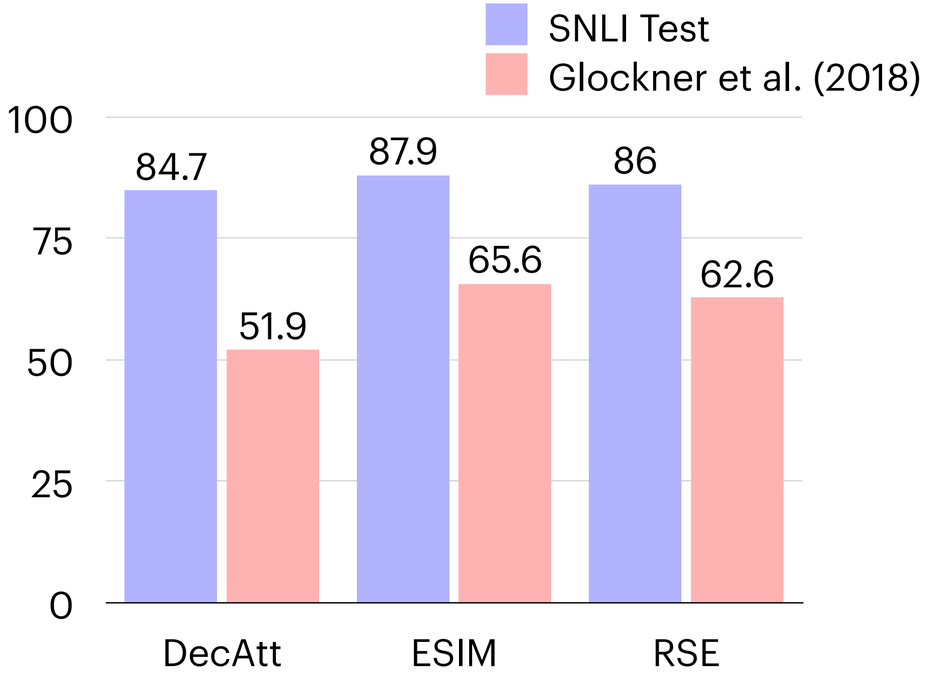
#2 Representations

2018

(ELMo-based Decomposable Attention)

P: Charlie will visit his mother in London on **Wednesday** evening.
H: Charlie will visit his mother in London on **Thursday** evening.

Prediction: Entailment (94.1%) ✘



Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. Max Glockner, **Vered Shwartz**, and Yoav Goldberg. ACL 2018

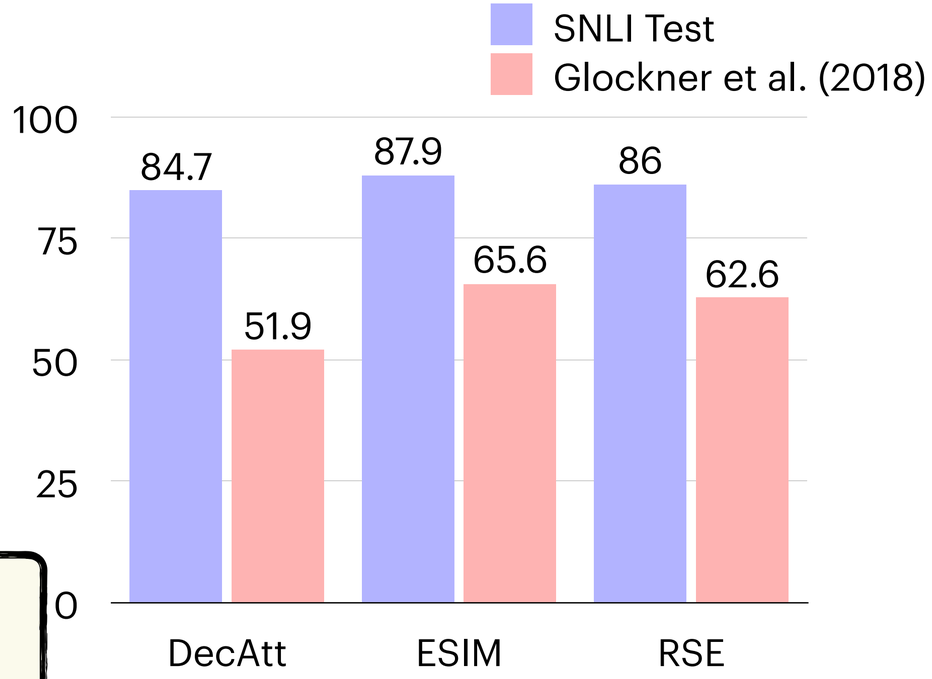
#2 Representations

2018

(ELMo-based Decomposable Attention)

P: Charlie will visit his mother in London on **Wednesday** evening.
H: Charlie will visit his mother in London on **Thursday** evening.

Prediction: Entailment (94.1%) ✘



- Errors: similar but mutually-exclusive words
- Accuracy increases with frequency in training set → Limited generalization ability!
- Similar findings in "NLI stress tests" (Naik et al., 2018)

Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. Max Glockner, **Vered Shwartz**, and Yoav Goldberg. ACL 2018

#2 Representations

2021

(RoBERTa)

https://demo.allennlp.org/textual-entailment

#2 Representations

2021

(RoBERTa)

P: Charlie will visit his mother in London on **Wednesday** evening.
H: Charlie will visit his mother in London on **Thursday** evening.

Prediction: Contradiction (73.2%) ✓

https://demo.allennlp.org/textual-entailment

#2 Representations

2021

(RoBERTa)

P: Charlie will visit his mother in London on **Wednesday** evening.

H: Charlie will visit his mother in London on **Thursday** evening.

Prediction: Contradiction (73.2%) ✓

P: Charlie said on **Wednesday** that he is busy on *Thursday* so he will visit his mother next week.

H: Charlie said on **Thursday** that he is busy on *Wednesday* so he will visit his mother next week.

Prediction: Entailment (92%) ✗

https://demo.allennlp.org/textual-entailment

#3 World Knowledge

> The definition of the textual entailment recognition task, like that of any other text understanding task, refers to human understanding of language. Such definition necessarily assumes common background knowledge, on which the (human) entailment judgment relies. [⋯] this knowledge should cover both extra-linguistic world knowledge [⋯] as well as knowledge of the language itself.

Recognizing Textual Entailment. Ido Dagan, Dan Roth, Mark Sammons, and Fabio Zanzotto. Morgan & Claypool Publishers, 2013.

#3 World Knowledge

Knowledge in Language Models

✓ Capture facts not explicitly mentioned in the corpus
(**Petroni et al. 2019; Feldman et al. 2019, this talk**)

#3 World Knowledge

Knowledge in Language Models

✓ Capture facts not explicitly mentioned in the corpus
(**Petroni et al. 2019; Feldman et al. 2019, this talk**)

✓ Associate concepts with their properties
(**Weir et al. 2020**)

Context	Human		RoBERTa-L	
	Response	PF	Response	p_{LM}
Everyone knows that a bear has ___	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02

#3 World Knowledge

Knowledge in Language Models

✓ Capture facts not explicitly mentioned in the corpus
(**Petroni et al. 2019; Feldman et al. 2019, this talk**)

✓ Associate concepts with their properties
(**Weir et al. 2020**)

Context	Human Response	Human PF	RoBERTa-L Response	RoBERTa-L p_{LM}
Everyone knows that a bear has ___	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02

#3 World Knowledge

Knowledge in Language Models

✗ Not sensitive to negation (**Kassner et al. 2020; Ettinger, 2020**)

✓ Capture facts not explicitly mentioned in the corpus (**Petroni et al. 2019; Feldman et al. 2019, this talk**)

✓ Associate concepts with their properties (**Weir et al. 2020**)

Context	Human		RoBERTa-L	
	Response	PF	Response	p_{LM}
Everyone knows that a bear has ___	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02

#3 World Knowledge

Knowledge in Language Models

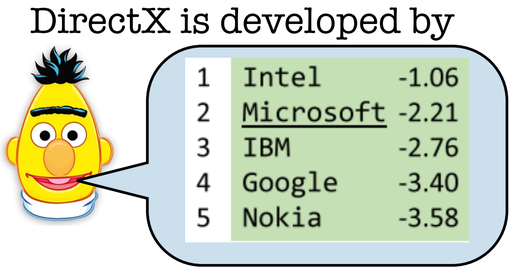
✓ Capture facts not explicitly mentioned in the corpus (**Petroni et al. 2019; Feldman et al. 2019, this talk**)

✓ Associate concepts with their properties (**Weir et al. 2020**)

Context	Human Response	Human PF	RoBERTa-L Response	RoBERTa-L p_{LM}
Everyone knows that a bear has ___	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02

✗ Not sensitive to negation (**Kassner et al. 2020; Ettinger, 2020**)

✗ Predict similar but mutually-exclusive facts (**Jiang et al., 2020**)



#3 World Knowledge

Knowledge in Language Models

✓ Capture facts not explicitly mentioned in the corpus (**Petroni et al. 2019; Feldman et al. 2019, this talk**)

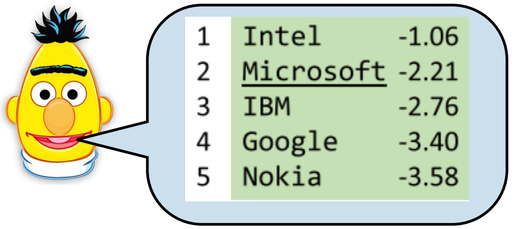
✓ Associate concepts with their properties (**Weir et al. 2020**)

Context	Human Response	Human PF	RoBERTa-L Response	RoBERTa-L p_{LM}
Everyone knows that a bear has ___	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02

✗ Not sensitive to negation (**Kassner et al. 2020; Ettinger, 2020**)

✗ Predict similar but mutually-exclusive facts (**Jiang et al., 2020**)

DirectX is developed by



✗ Lack perceptual and physical knowledge (**Forbes et al. 2019, Weir et al., 2020, Bisk et al. 2020**)

#3 World Knowledge

Knowledge in Language Models

✓ Capture facts not explicitly mentioned in the corpus (**Petroni et al. 2019; Feldman et al. 2019, this talk**)

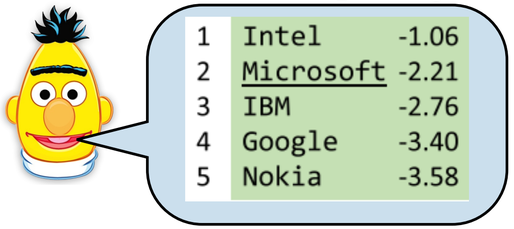
✓ Associate concepts with their properties (**Weir et al. 2020**)

Context	Human Response	Human PF	RoBERTa-L Response	RoBERTa-L p_{LM}
Everyone knows that a bear has ___	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02

✗ Not sensitive to negation (**Kassner et al. 2020; Ettinger, 2020**)

✗ Predict similar but mutually-exclusive facts (**Jiang et al., 2020**)

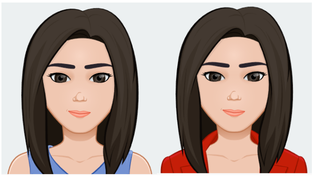
DirectX is developed by



✗ Lack perceptual and physical knowledge (**Forbes et al. 2019, Weir et al., 2020, Bisk et al. 2020**)

✗ Don't differentiate constant vs. contingent facts

Zebras are black and white.



My shirt is blue / red.

#3 World Knowledge

Knowledge in Language Models

✓ Capture facts not explicitly mentioned in the corpus (**Petroni et al. 2019; Feldman et al. 2019, this talk**)

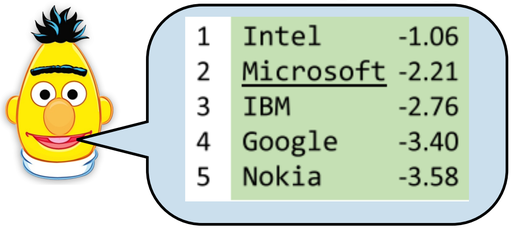
✓ Associate concepts with their properties (**Weir et al. 2020**)

Context	Human Response	Human PF	RoBERTa-L Response	RoBERTa-L p_{LM}
Everyone knows that a bear has ___	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02

✗ Not sensitive to negation (**Kassner et al. 2020; Ettinger, 2020**)

✗ Predict similar but mutually-exclusive facts (**Jiang et al., 2020**)

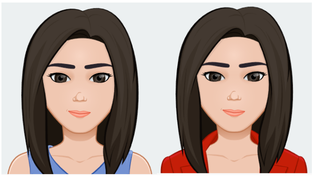
DirectX is developed by



✗ Lack perceptual and physical knowledge (**Forbes et al. 2019, Weir et al., 2020, Bisk et al. 2020**)

✗ Don't differentiate constant vs. contingent facts

Zebras are black and white.



My shirt is blue / red.

✗ Reporting bias

Do Neural Language Models Overcome Reporting Bias? **Vered Shwartz** and Yejin Choi. COLING 2020.

#3 World Knowledge

Knowledge in Language Models

✓ Capture facts not explicitly mentioned in the corpus (**Petroni et al. 2019; Feldman et al. 2019, this talk**)

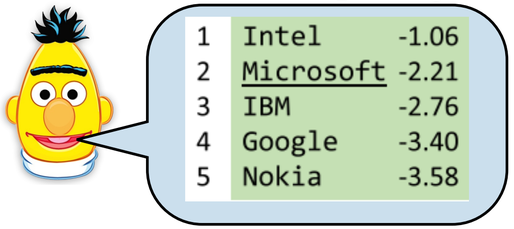
✓ Associate concepts with their properties (**Weir et al. 2020**)

Context	Human Response	Human PF	RoBERTa-L Response	RoBERTa-L p_{LM}
Everyone knows that a bear has ___	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02

✗ Not sensitive to negation (**Kassner et al. 2020; Ettinger, 2020**)

✗ Predict similar but mutually-exclusive facts (**Jiang et al., 2020**)

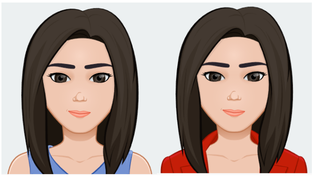
DirectX is developed by



✗ Lack perceptual and physical knowledge (**Forbes et al. 2019, Weir et al., 2020, Bisk et al. 2020**)

✗ Don't differentiate constant vs. contingent facts

Zebras are black and white.



My shirt is blue / red.

✗ Reporting bias — The man turned on the faucet. As a result,

Do Neural Language Models Overcome Reporting Bias? **Vered Shwartz** and Yejin Choi. COLING 2020.

#3 World Knowledge

Knowledge in Language Models

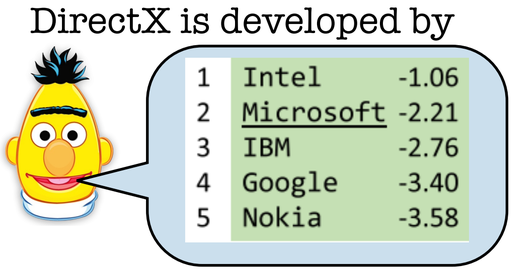
✓ Capture facts not explicitly mentioned in the corpus (**Petroni et al. 2019; Feldman et al. 2019, this talk**)

✓ Associate concepts with their properties (**Weir et al. 2020**)

Context	Human Response	Human PF	RoBERTa-L Response	RoBERTa-L p_{LM}
Everyone knows that a bear has ___	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02

✗ Not sensitive to negation (**Kassner et al. 2020; Ettinger, 2020**)

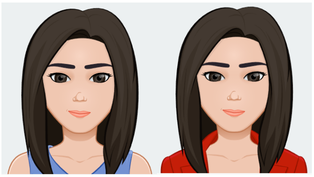
✗ Predict similar but mutually-exclusive facts (**Jiang et al., 2020**)



✗ Lack perceptual and physical knowledge (**Forbes et al. 2019, Weir et al., 2020, Bisk et al. 2020**)

✗ Don't differentiate constant vs. contingent facts

Zebras are black and white.



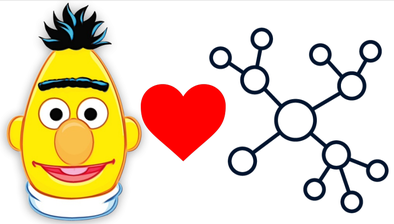
My shirt is blue / red.

✗ Reporting bias

The man turned on the faucet. As a result, **the man's blood was sprayed everywhere.**

Do Neural Language Models Overcome Reporting Bias? **Vered Shwartz** and Yejin Choi. COLING 2020.

Limitations of neural NLI Models



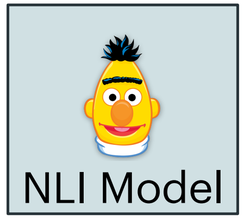
Incorporating symbolic knowledge into neural NLI models

NLI is too easy? What's next?

Learning Semantic Phenomena

The office is located in Baytown [SEP]
The office is located in Texas

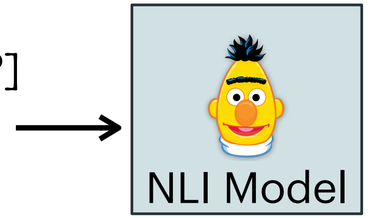
label: entailment



Learning Semantic Phenomena

The office is located in Baytown [SEP]
The office is located in Texas

label: entailment



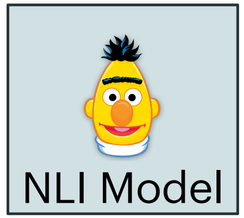
Model expected to learn:

Fact: LocatedIn(Baytown, Texas)

Learning Semantic Phenomena

The office is located in Baytown [SEP]
The office is located in Texas

label: entailment

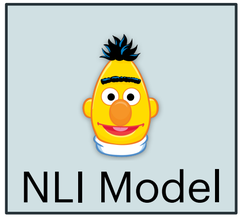


Model expected to learn:

Fact: LocatedIn(Baytown, Texas)

Rule: Entailment between a city and its state in upward monotone sentences

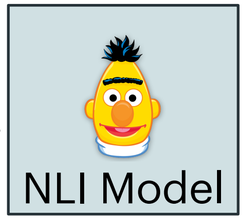
Probing and Inoculation



Probing and Inoculation

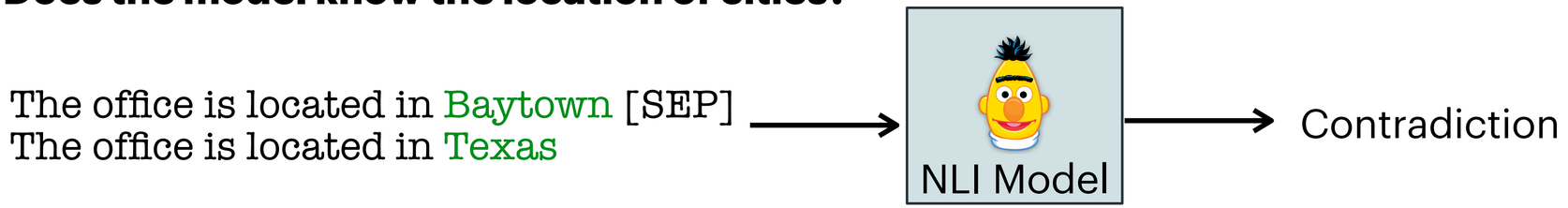
Does the model know the location of cities?

The office is located in Baytown [SEP]
The office is located in Texas



Probing and Inoculation

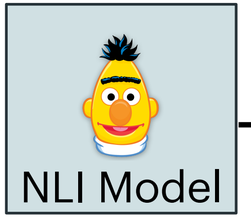
Does the model know the location of cities?



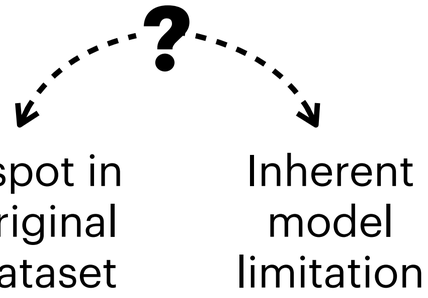
Probing and Inoculation

Does the model know the location of cities?

The office is located in Baytown [SEP]
The office is located in Texas



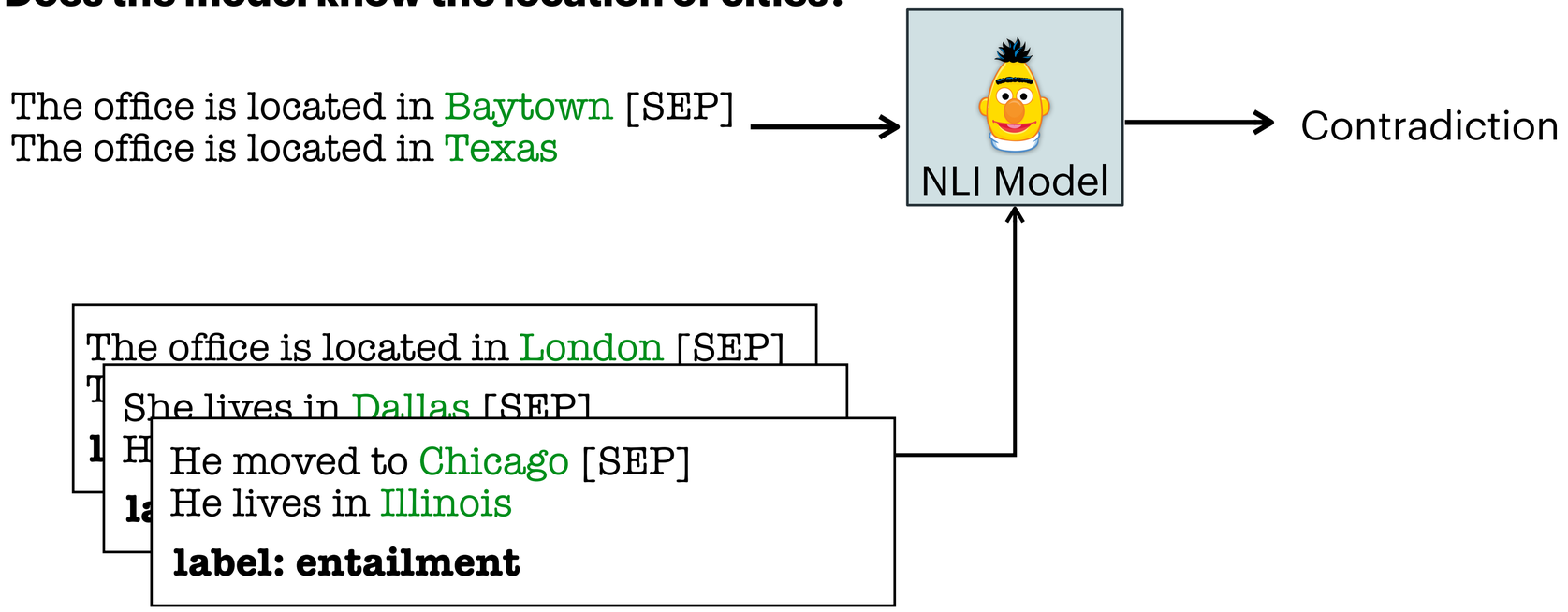
Contradiction

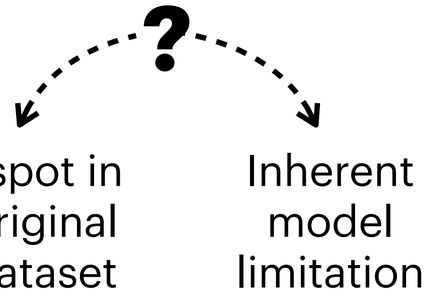


Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. Nelson F. Liu, Roy Schwartz, and Noah A. Smith. NAACL 2019.

Probing and Inoculation

Does the model know the location of cities?

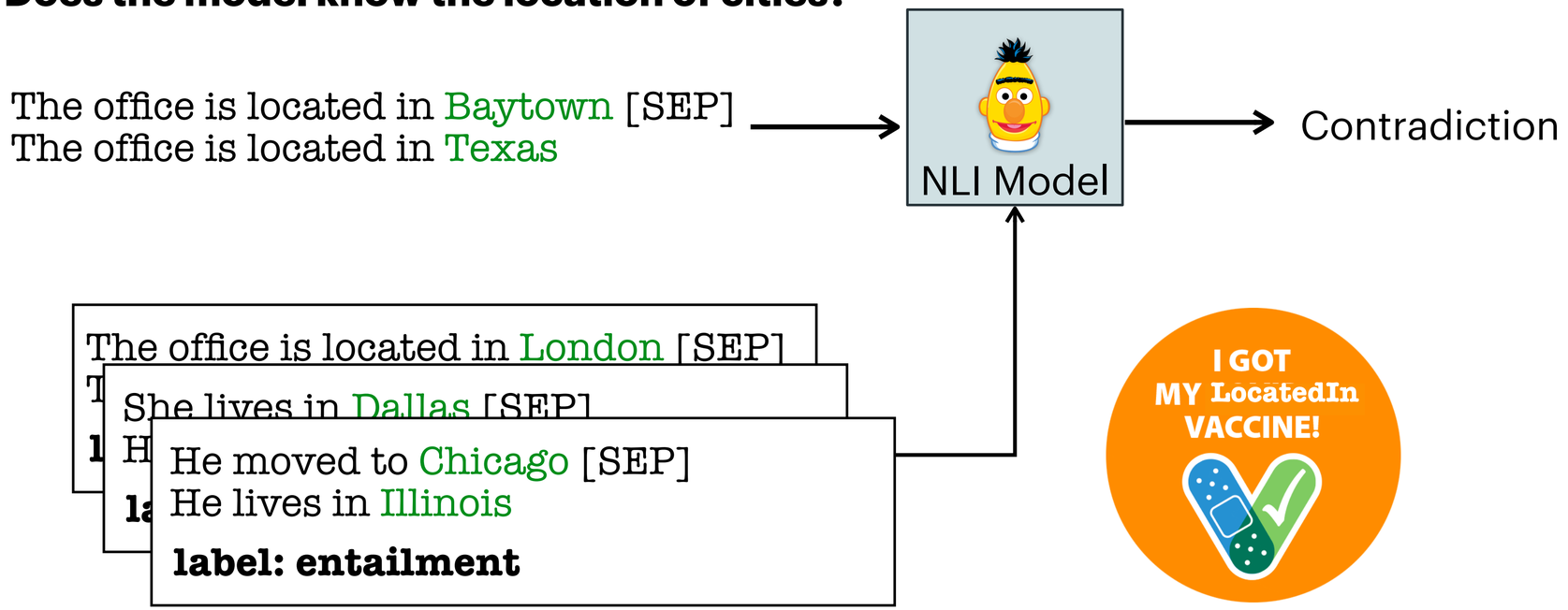


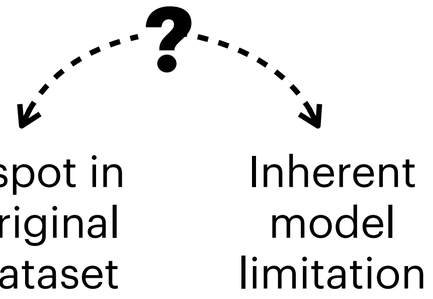


Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. Nelson F. Liu, Roy Schwartz, and Noah A. Smith. NAACL 2019.

Probing and Inoculation

Does the model know the location of cities?

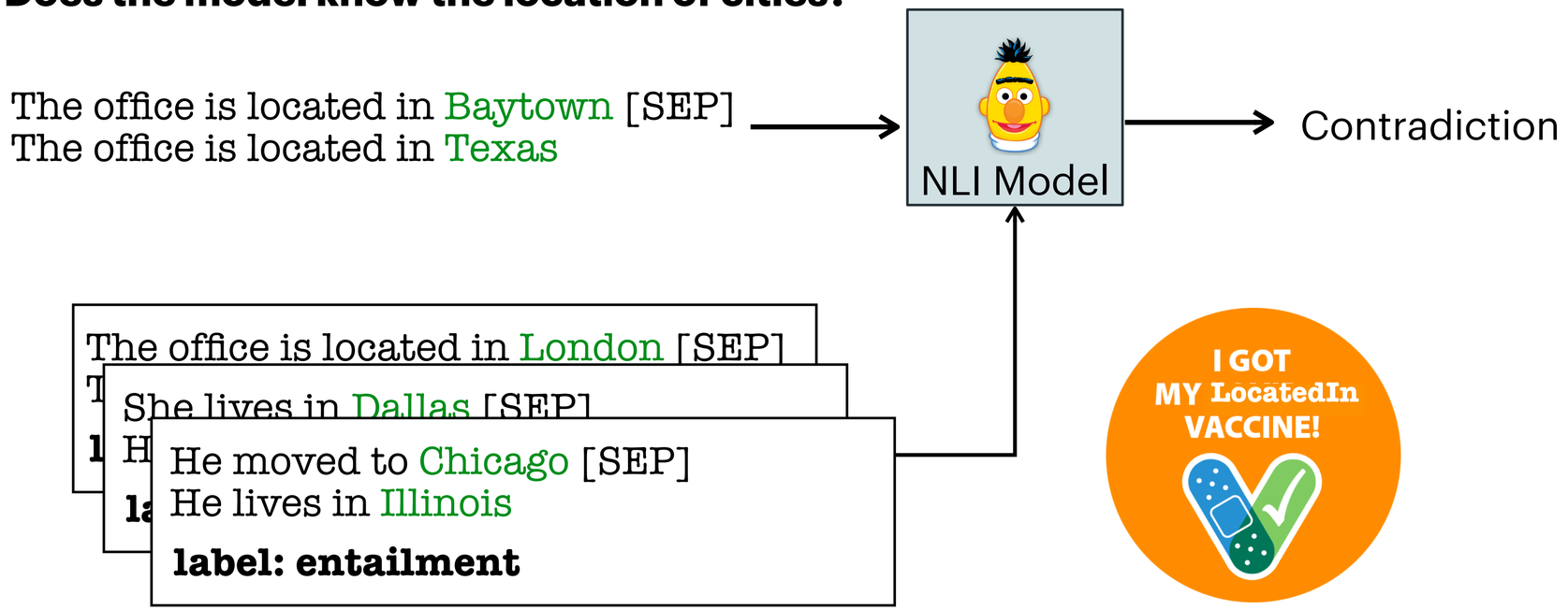




Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. Nelson F. Liu, Roy Schwartz, and Noah A. Smith. NAACL 2019.

Probing and Inoculation

Does the model know the location of cities?



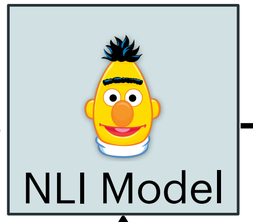
Possible Outcomes

Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. Nelson F. Liu, Roy Schwartz, and Noah A. Smith. NAACL 2019.

Probing and Inoculation

Does the model know the location of cities?

The office is located in Baytown [SEP]
The office is located in Texas



Entailment

The office is located in London [SEP]
She lives in Dallas [SEP]
He moved to Chicago [SEP]
He lives in Illinois
label: entailment

Possible Outcomes

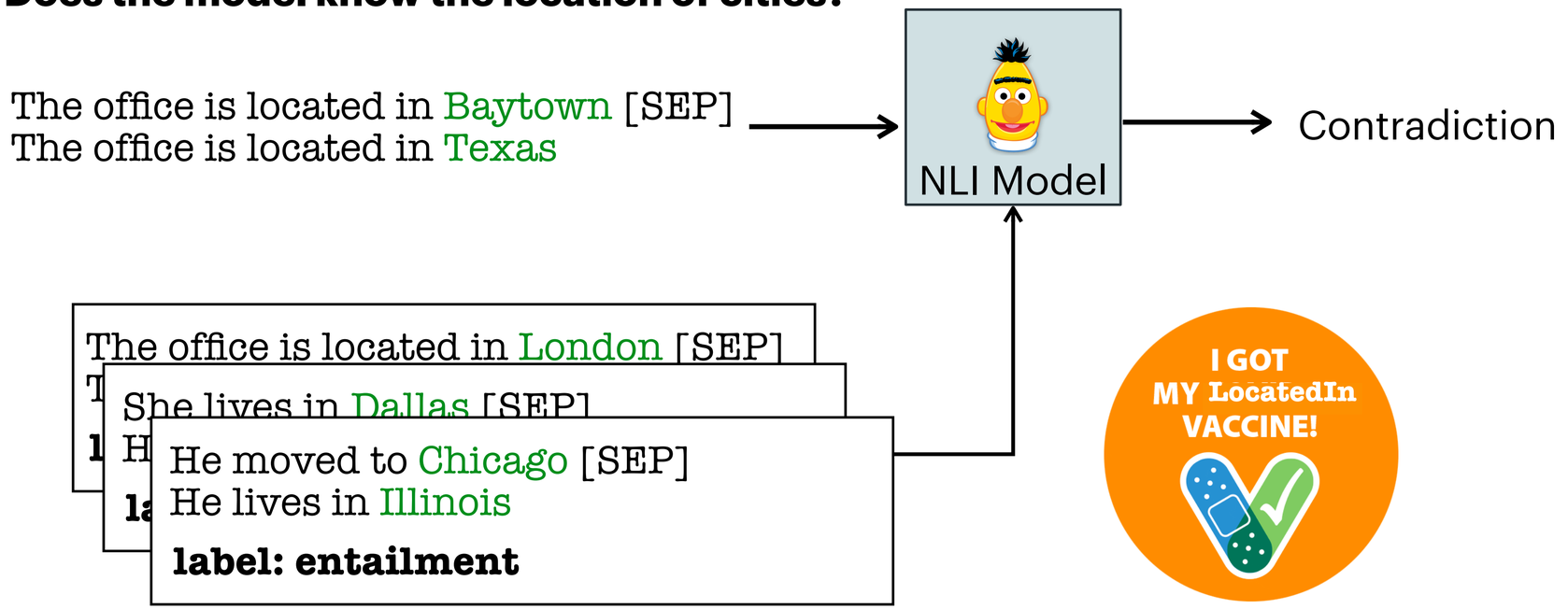
✓Success

Blindspot in the original NLI dataset

Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. Nelson F. Liu, Roy Schwartz, and Noah A. Smith. NAACL 2019.

Probing and Inoculation

Does the model know the location of cities?

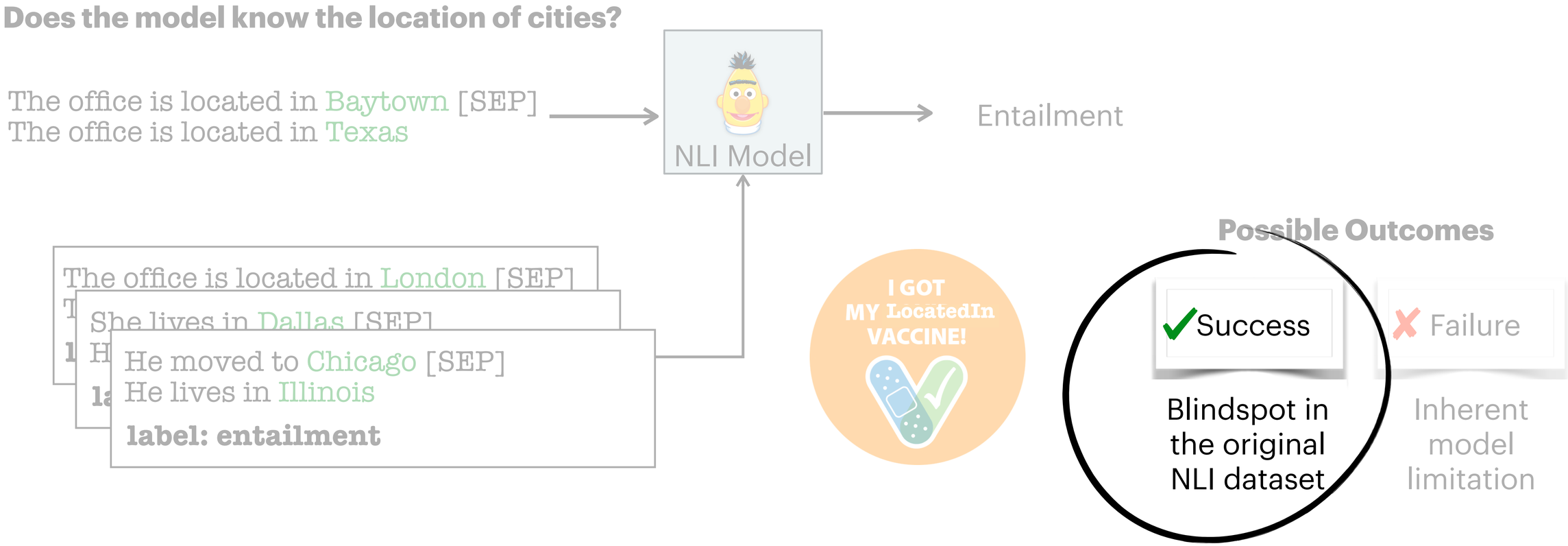


Possible Outcomes

✓ Success	✗ Failure
Blindspot in the original NLI dataset	Inherent model limitation

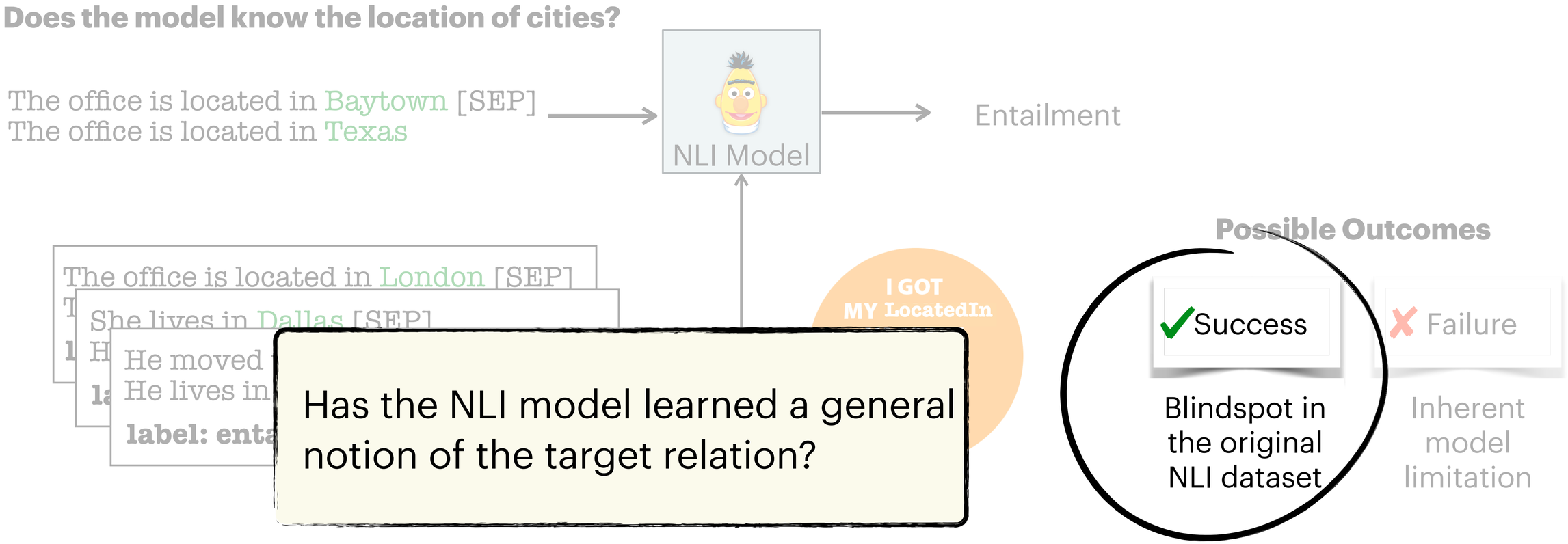
Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. Nelson F. Liu, Roy Schwartz, and Noah A. Smith. NAACL 2019.

Probing and Inoculation



Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. Nelson F. Liu, Roy Schwartz, and Noah A. Smith. NAACL 2019.

Probing and Inoculation



Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. Nelson F. Liu, Roy Schwartz, and Noah A. Smith. NAACL 2019.

Analyzing Generalization via Controlled Variance

Numerical Reasoning

P: I see 260 coins in the bucket.

H: I see more than 232 coins in the bucket.

Label: Entailment

Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets Ohad Rozen, **Vered Shwartz**, Roee Aharoni, and Ido Dagan. CoNLL 2019.

Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus

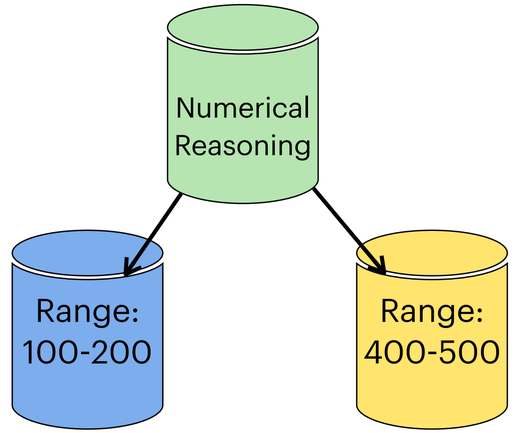
Numerical Reasoning

P: I see 260 coins in the bucket.

H: I see more than 232 coins in the bucket.

Label: Entailment

Axis: number range



Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets Ohad Rozen, **Vered Shwartz**, Roee Aharoni, and Ido Dagan. CoNLL 2019.

Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus
2. Fine-tune on one set and test on another

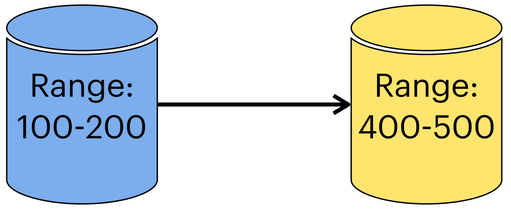
Numerical Reasoning

P: I see 260 coins in the bucket.

H: I see more than 232 coins in the bucket.

Label: Entailment

Axis: number range



Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets Ohad Rozen, **Vered Shwartz**, Roee Aharoni, and Ido Dagan. CoNLL 2019.

Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus
2. Fine-tune on one set and test on another

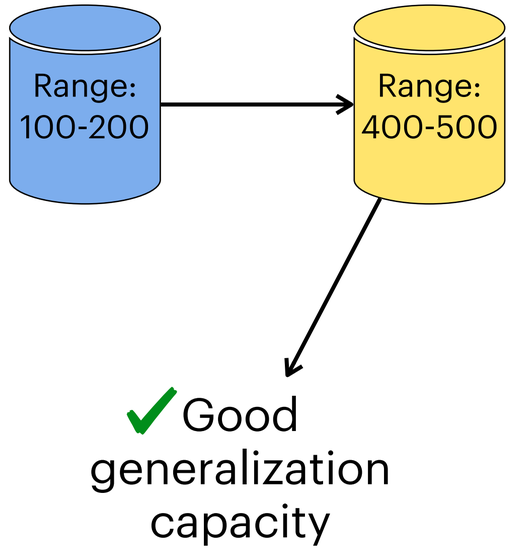
Numerical Reasoning

P: I see 260 coins in the bucket.

H: I see more than 232 coins in the bucket.

Label: Entailment

Axis: number range



Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets Ohad Rozen, **Vered Shwartz**, Roee Aharoni, and Ido Dagan. CoNLL 2019.

Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus
2. Fine-tune on one set and test on another

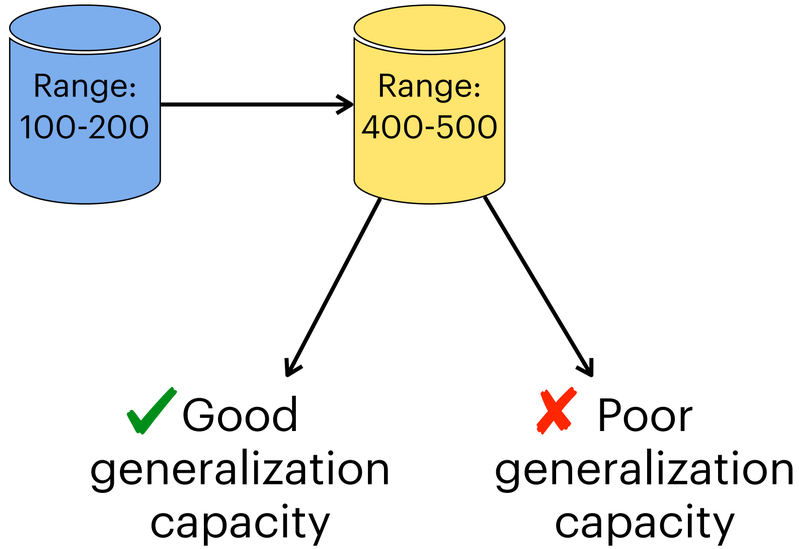
Numerical Reasoning

P: I see 260 coins in the bucket.

H: I see more than 232 coins in the bucket.

Label: Entailment

Axis: number range



Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets Ohad Rozen, **Vered Shwartz**, Roee Aharoni, and Ido Dagan. CoNLL 2019.

Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus
2. Fine-tune on one set and test on another

Numerical Reasoning

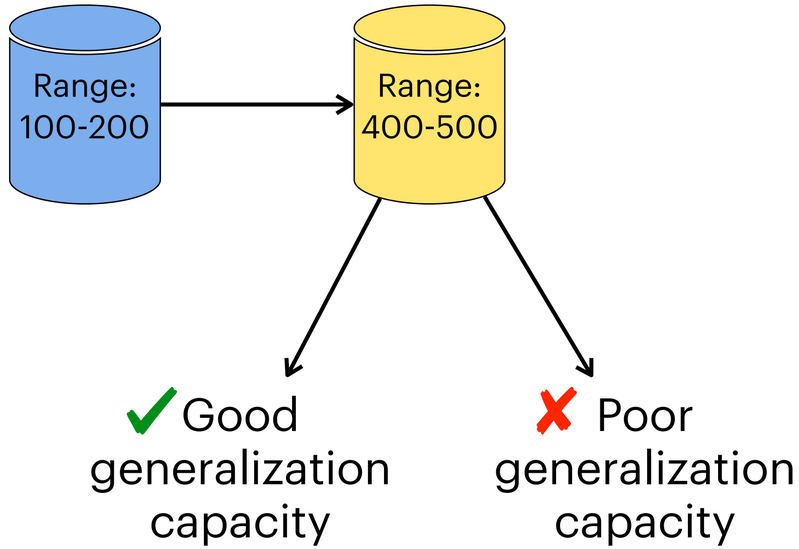
P: I see 260 coins in the bucket.

H: I see more than 232 coins in the bucket.

Label: Entailment

✗ Axis: number range

✓ Axis: syntactic complexity



Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets Ohad Rozen, **Vered Shwartz**, Roee Aharoni, and Ido Dagan. CoNLL 2019.

Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus
2. Fine-tune on one set and test on another

Numerical Reasoning

P: I see 260 coins in the bucket.

H: I see more than 232 coins in the bucket.

Label: Entailment

Dative Alternation

P: I baked my mom a cake.

H: I baked a cake for my mom.

Label: Entailment

✓ Axis: lexical variability

✗ Axis: syntactic complexity

Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets Ohad Rozen, **Vered Shwartz**, Roee Aharoni, and Ido Dagan. CoNLL 2019.

Analyzing Generalization via Controlled Variance

1. Train/dev/test split across the dimension in focus
2. Fine-tune on one set and test on another

Numerical Reasoning

P: I see 260 coins in the bucket.

H: I see more than 232 coins in the bucket.

Label: Entailment

Dative Alternation

P: I baked my mom a cake.

H: I baked a cake for my mom.

Label: Entailment

✓ Axis: lexical variability

✗ Axis: syntactic complexity

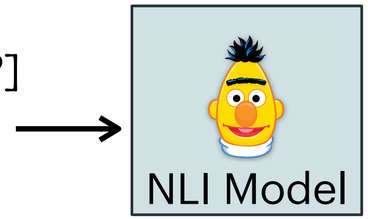
May decrease performance on the main task (Richardson et al., 2020)

Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets Ohad Rozen, **Vered Shwartz**, Roee Aharoni, and Ido Dagan. CoNLL 2019.

Incorporating Knowledge into NLI Models

The office is located in Baytown [SEP]
The office is located in Texas

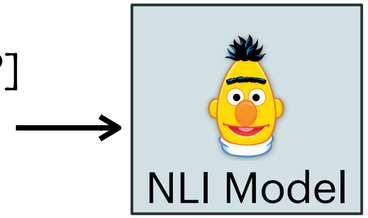
label: entailment



Incorporating Knowledge into NLI Models

The office is located in Baytown [SEP]
The office is located in Texas

label: entailment



Model learns:

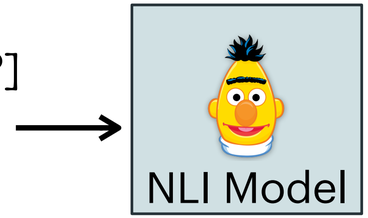
Fact: LocatedIn(Baytown, Texas)

Rule: Entailment between a city and its state in upward monotone sentences

Incorporating Knowledge into NLI Models

The office is located in Baytown [SEP]
The office is located in Texas

label: entailment



Model learns:

Fact: LocatedIn(Baytown, Texas)

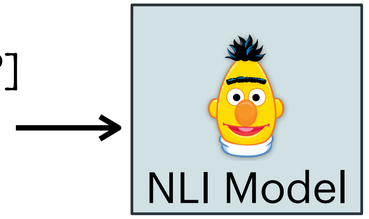
Rule: Entailment between a city and its state in upward monotone sentences

May be learned given enough data.

Incorporating Knowledge into NLI Models

The office is located in Baytown [SEP]
The office is located in Texas

label: entailment



Model learns:

Fact: LocatedIn(Baytown, Texas)

Rule: Entailment between a city and its state in upward monotone sentences

May be learned given enough data.

Impossible (and inefficient) to teach
an NLI model every fact it might need.

Incorporating Knowledge into NLI Models

The office is located in Baytown [SEP]
The office is located in Texas

label: entailment

WIKIPEDIA
The Free Encyclopedia

Fact: LocatedIn(Baytown, Texas)

Knowledge Informed NLI Model

Model learns:

Rule: Entailment between a city and its state in upward monotone sentences

Incorporating Knowledge into NLI Models

The office is located in Baytown [SEP]
The office is located in Texas

label: entailment

WIKIPEDIA
The Free Encyclopedia

Fact: LocatedIn(Baytown, Texas)

Knowledge Informed NLI Model

Model learns:

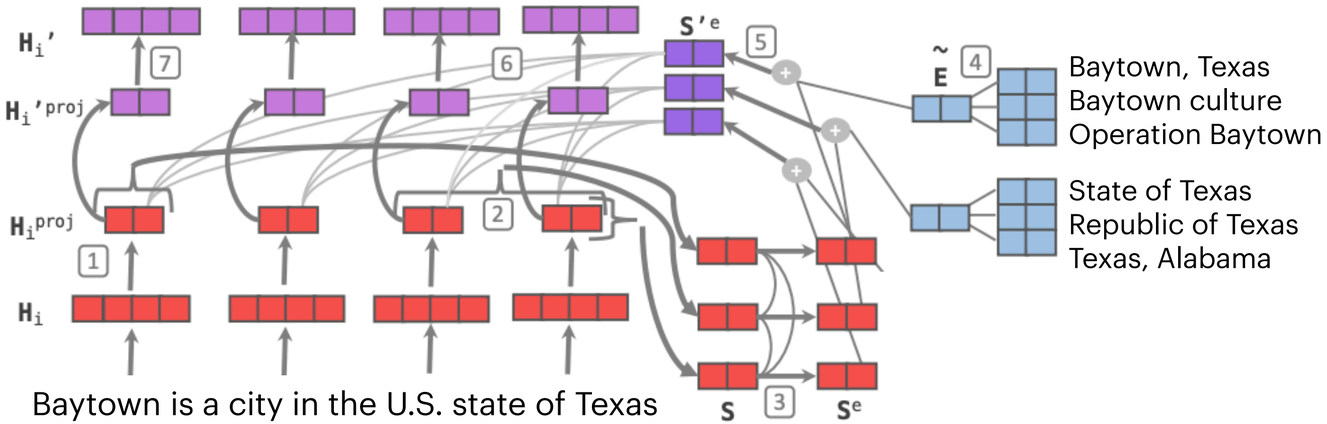
Rule: Entailment between a city and its state in upward monotone sentences

How to incorporate relational knowledge?

Incorporating Factual Knowledge into LMs

Entity-Centric Knowledge

KnowBERT

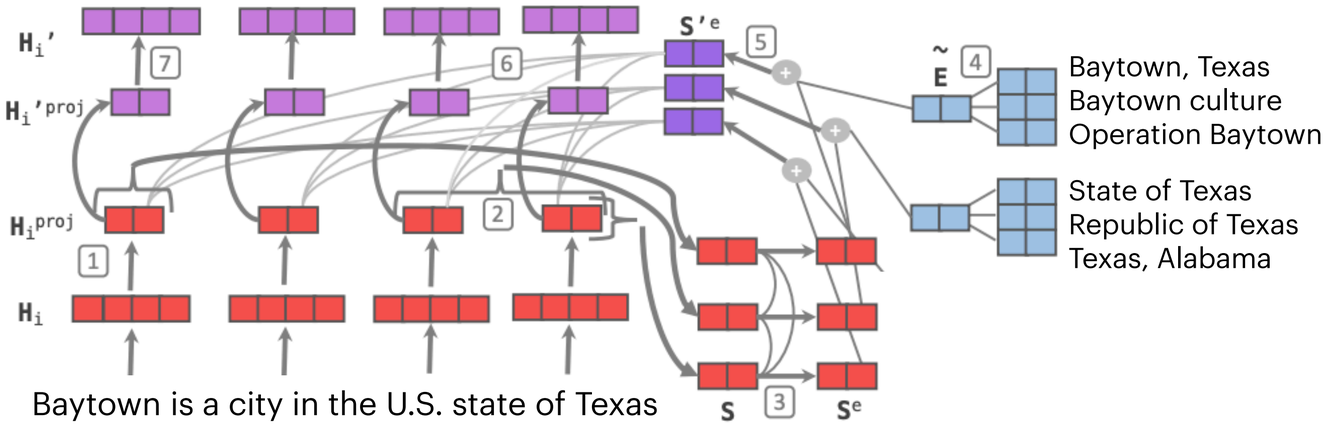


Knowledge Enhanced Contextual Word Representations. Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. EMNLP 2019.

Incorporating Factual Knowledge into LMs

Entity-Centric Knowledge

KnowBERT



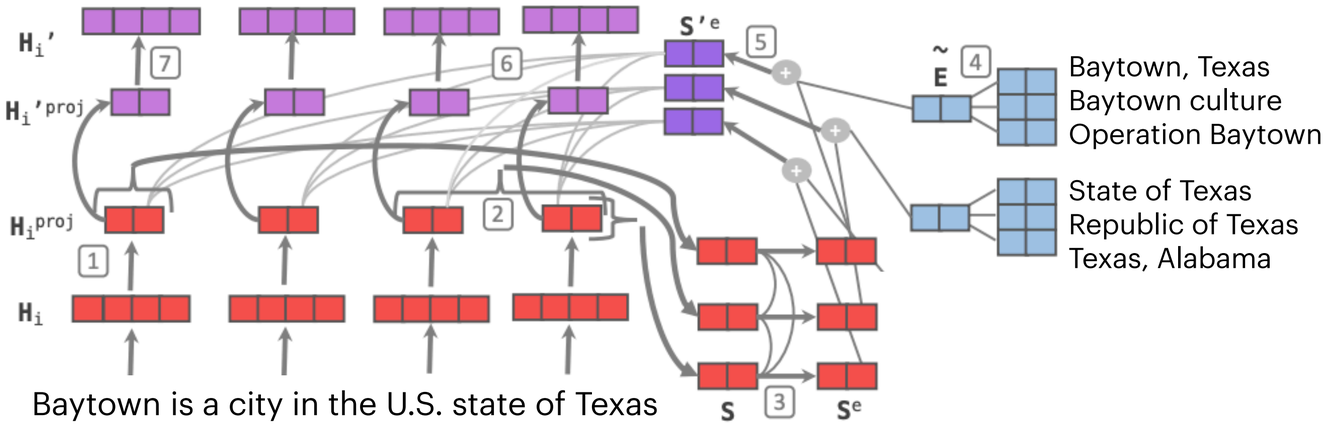
😃 performance improvement on relation extraction, entity typing, WSD

Knowledge Enhanced Contextual Word Representations. Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. EMNLP 2019.

Incorporating Factual Knowledge into LMs

Entity-Centric Knowledge

KnowBERT



😢 can't learn new facts after training

The office is located in Nanhui [SEP]
The office is located in China

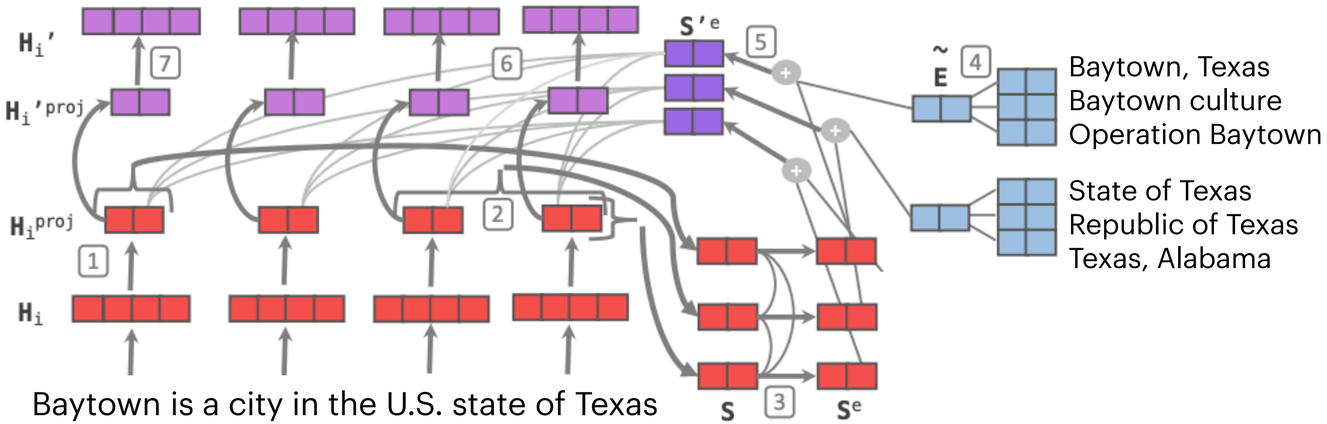
😃 performance improvement on relation extraction, entity typing, WSD

Knowledge Enhanced Contextual Word Representations. Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. EMNLP 2019.

Incorporating Factual Knowledge into LMs

Entity-Centric Knowledge

KnowBERT



😃 performance improvement on relation extraction, entity typing, WSD

😢 can't learn new facts after training

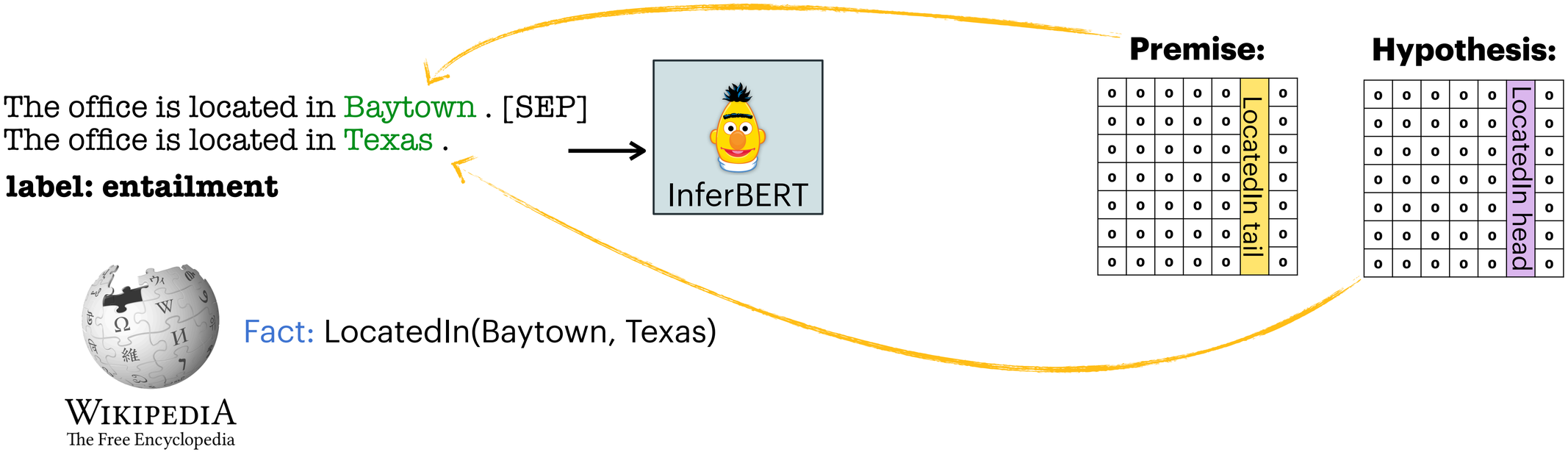
The office is located in Nanhui [SEP]
The office is located in China

😐 computationally expensive to re-train

Knowledge Enhanced Contextual Word Representations. Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. EMNLP 2019.

Incorporating Factual Knowledge into LMs

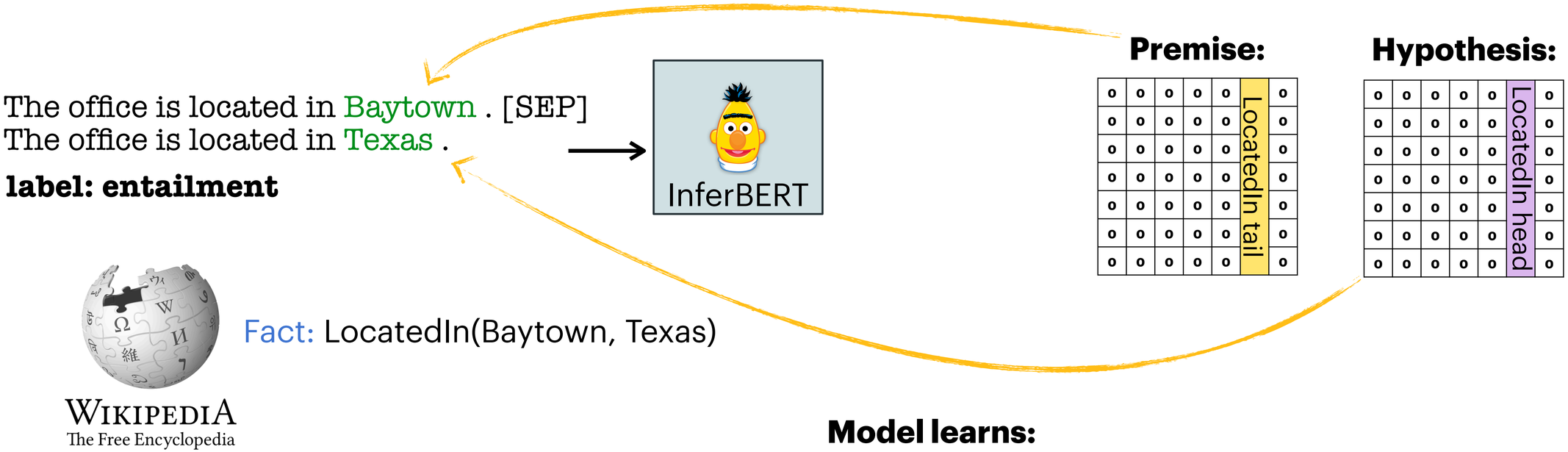
Relation-Centric Knowledge



Teach the Rules, Provide the Facts: Targeted Relational-knowledge Enhancement for Textual Inference. Ohad Rozen, Shmuel Amar, **Vered Shwartz** and Ido Dagan. *SEM 2021

Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge



Model learns:

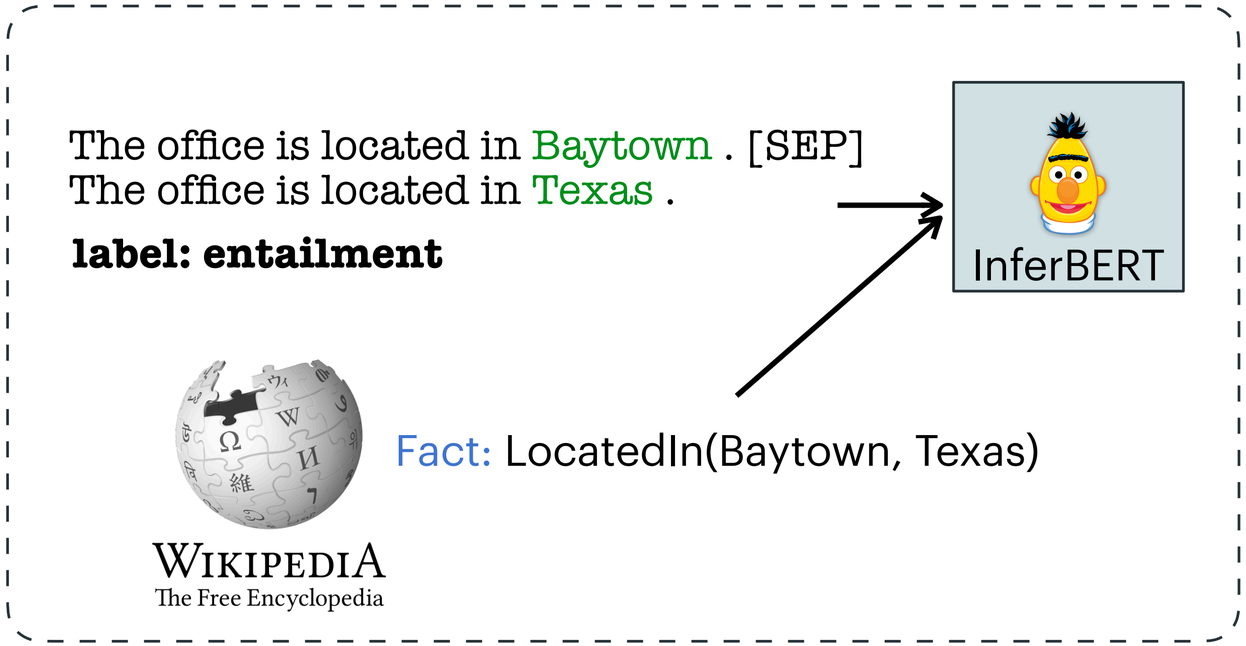
Rule: Entailment between a city and its state in upward monotone sentences

Relation Embeddings

Teach the Rules, Provide the Facts: Targeted Relational-knowledge Enhancement for Textual Inference. Ohad Rozen, Shmuel Amar, **Vered Shwartz** and Ido Dagan. *SEM 2021

Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge



< 700 training examples

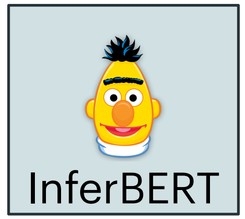
Teach the Rules, Provide the Facts: Targeted Relational-knowledge Enhancement for Textual Inference. Ohad Rozen, Shmuel Amar, **Vered Shwartz** and Ido Dagan. *SEM 2021

Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge

The office is located in Baytown . [SEP]
The office is located in Texas .

label: entailment



WIKIPEDIA
The Free Encyclopedia

Fact: LocatedIn(Baytown, Texas)

< 700 training examples

😃 performance improvement on NLI challenge sets

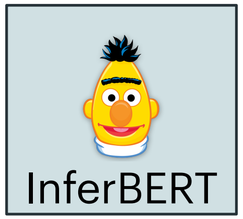
Teach the Rules, Provide the Facts: Targeted Relational-knowledge Enhancement for Textual Inference. Ohad Rozen, Shmuel Amar, **Vered Shwartz** and Ido Dagan. *SEM 2021

Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge

The office is located in Baytown . [SEP]
The office is located in Texas .

label: entailment



WIKIPEDIA
The Free Encyclopedia

Fact: LocatedIn(Baytown, Texas)

< 700 training examples

😃 performance improvement on NLI challenge sets

😃 can retrieve and use facts about unseen entities

The office is located in Nanhui [SEP]
The office is located in China

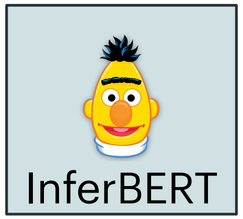
Teach the Rules, Provide the Facts: Targeted Relational-knowledge Enhancement for Textual Inference. Ohad Rozen, Shmuel Amar, **Vered Shwartz** and Ido Dagan. *SEM 2021

Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge

The office is located in Baytown . [SEP]
The office is located in Texas .

label: entailment



WIKIPEDIA
The Free Encyclopedia

Fact: LocatedIn(Baytown, Texas)

< 700 training examples

😃 performance improvement on NLI challenge sets

😃 can retrieve and use facts about unseen entities

The office is located in Nanhui [SEP]
The office is located in China

😃 computationally efficient

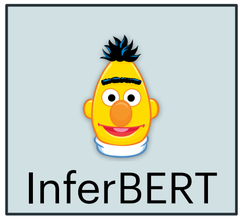
Teach the Rules, Provide the Facts: Targeted Relational-knowledge Enhancement for Textual Inference. Ohad Rozen, Shmuel Amar, **Vered Shwartz** and Ido Dagan. *SEM 2021

Incorporating Factual Knowledge into LMs

Relation-Centric Knowledge

The office is located in Baytown . [SEP]
The office is located in Texas .

label: entailment



WIKIPEDIA
The Free Encyclopedia

Fact: LocatedIn(Baytown, Texas)

< 700 training examples

😃 performance improvement on NLI challenge sets

😃 can retrieve and use facts about unseen entities

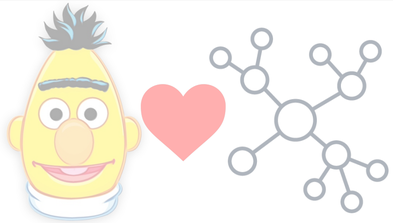
The office is located in Nanhui [SEP]
The office is located in China

😃 computationally efficient

😐 task-specific knowledge incorporation

Teach the Rules, Provide the Facts: Targeted Relational-knowledge Enhancement for Textual Inference. Ohad Rozen, Shmuel Amar, **Vered Shwartz** and Ido Dagan. *SEM 2021

Limitations of neural NLI Models



Incorporating symbolic knowledge into neural NLI models

NLI is too easy? What's next?

Real-World NLI #1

Partial Entailment

S_1: Amazon to acquire Whole Foods Market for $13.7 Billion.

S_2: Amazon is buying Whole Foods Market for almost $14 Billion in cash.

Real-World NLI #1

Partial Entailment

S_1: Amazon to acquire Whole Foods Market for $13.7 Billion.

S_2: Amazon is buying Whole Foods Market for almost $14 Billion in cash.

Real-World NLI #1

Partial Entailment

S_1: Amazon to acquire Whole Foods Market for $13.7 Billion.

S_2: Amazon is buying Whole Foods Market for almost $14 Billion in cash.

Subjectivity and inherent disagreements (Pavlick and Kwiatkowski, 2019)

Real-World NLI #2

Defeasible Natural Language Inference

P: Tweety is a bird.

H: Tweety flies.

Real-World NLI #2

Defeasible Natural Language Inference

P: Tweety is a bird.

H: Tweety flies.

NLI model: Entailment

Real-World NLI #2

Defeasible Natural Language Inference

P: Tweety is a bird.

H: Tweety flies.

NLI model: Entailment

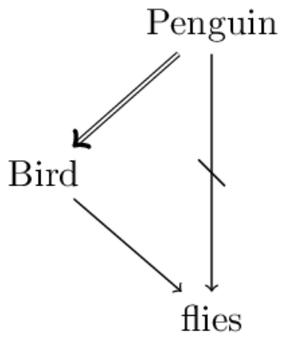
Skeptical NLI model: given the information I currently have, I suppose so, but I can think of cases in which this is false.

Defeasible Inference (Reiter, 1980)

Given premise P, a hypothesis H is **defeasible** if there exists an update U (consistent with P) such that a human would find H less likely to be true after learning U.

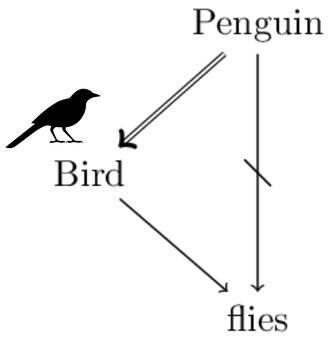
Defeasible Inference (Reiter, 1980)

Given premise P, a hypothesis H is **defeasible** if there exists an update U (consistent with P) such that a human would find H less likely to be true after learning U.



Defeasible Inference (Reiter, 1980)

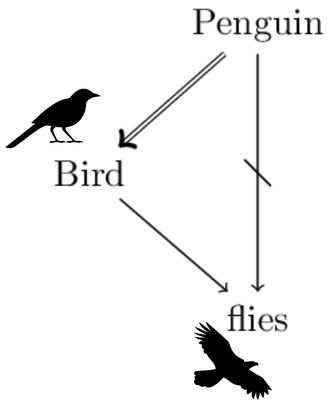
Given premise P, a hypothesis H is **defeasible** if there exists an update U (consistent with P) such that a human would find H less likely to be true after learning U.



P: Tweety is a bird.

Defeasible Inference (Reiter, 1980)

Given premise P, a hypothesis H is **defeasible** if there exists an update U (consistent with P) such that a human would find H less likely to be true after learning U.

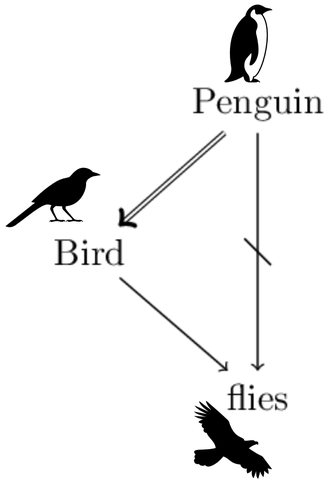


P: Tweety is a bird.

H: Tweety flies.

Defeasible Inference (Reiter, 1980)

Given premise P, a hypothesis H is **defeasible** if there exists an update U (consistent with P) such that a human would find H less likely to be true after learning U.



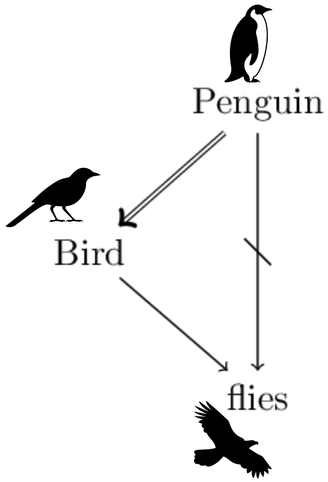
P: Tweety is a bird.

H: Tweety flies.

U: Tweety is a penguin.

Defeasible Inference (Reiter, 1980)

Given premise P, a hypothesis H is **defeasible** if there exists an update U (consistent with P) such that a human would find H less likely to be true after learning U.



P: Tweety is a bird.

H: Tweety flies.

U: Tweety is a penguin.

Useful for **Real-time Summarization**: Facts change as the story unfolds.

Defeasible Inference in Natural Language

An update U is called a **weakener** if, given a premise P and hypothesis H, a human would most likely find H *less likely to be true* after learning U; if they would find H *more likely to be true*, then we call U a **strengthener**.

P: Tweety is a bird. 

H: Tweety flies. 

Weakener: Tweety is a penguin. 

Thinking Like a Skeptic: Defeasible Inference in Natural Language.

Rachel Rudinger, **Vered Shwartz**, Jena Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah Smith, and Yejin Choi. Findings of EMNLP 2020.

Defeasible Inference in Natural Language

An update U is called a **weakener** if, given a premise P and hypothesis H, a human would most likely find H *less likely to be true* after learning U; if they would find H *more likely to be true*, then we call U a **strengthener**.

P: Tweety is a bird. 

H: Tweety flies. 

Weakener: Tweety is a penguin. 

Strengthener: Tweety is on a tree. 

Thinking Like a Skeptic: Defeasible Inference in Natural Language.

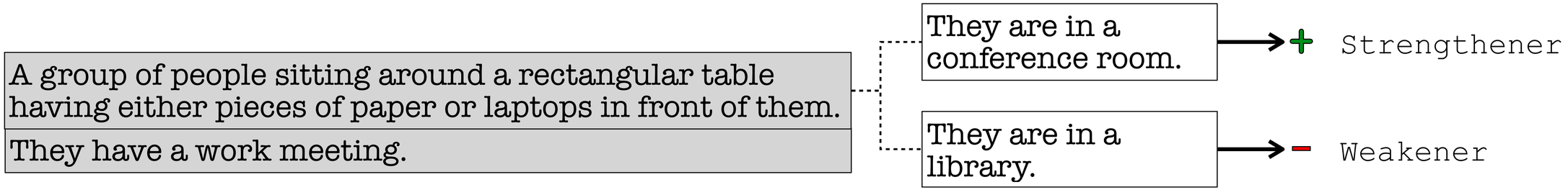
Rachel Rudinger, **Vered Shwartz**, Jena Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah Smith, and Yejin Choi. Findings of EMNLP 2020.

Defeasible Inference in Natural Language

Defeasible Inference in Natural Language

Discriminative Task

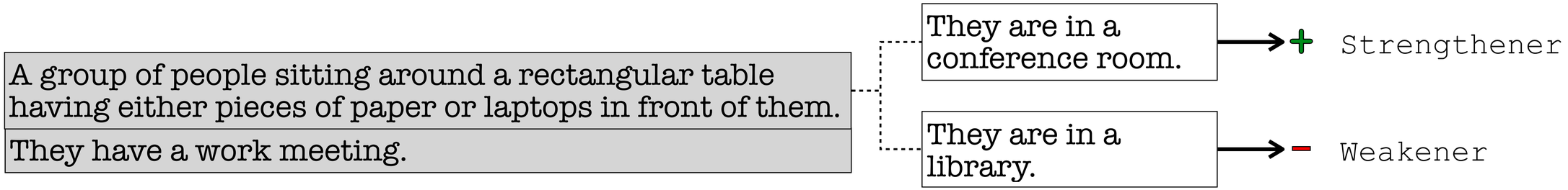
Determine whether an update weakens or strengthens the hypothesis.



Defeasible Inference in Natural Language

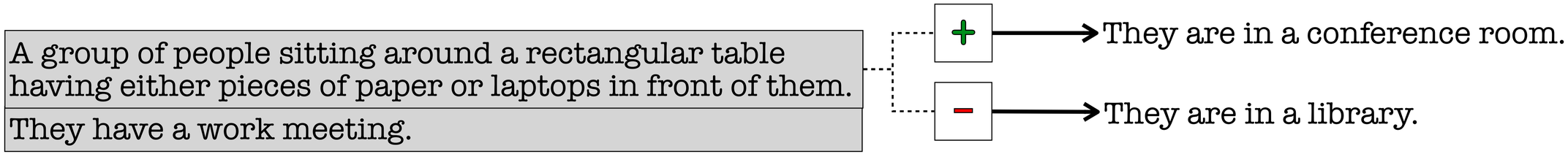
Discriminative Task

Determine whether an update weakens or strengthens the hypothesis.



Generative Task

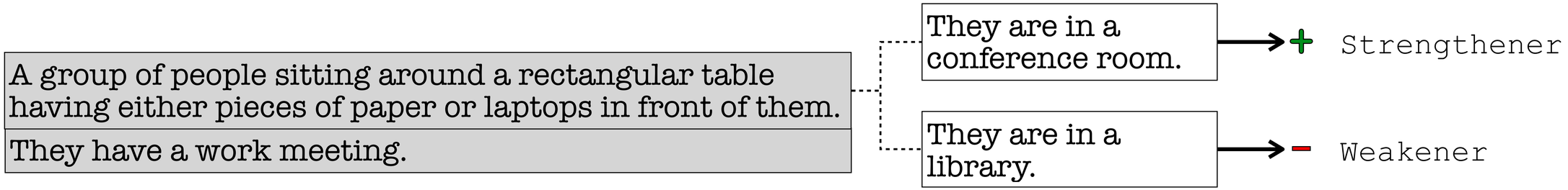
Generate a weakening or strengthening update for a given premise-hypothesis pair.



Defeasible Inference in Natural Language

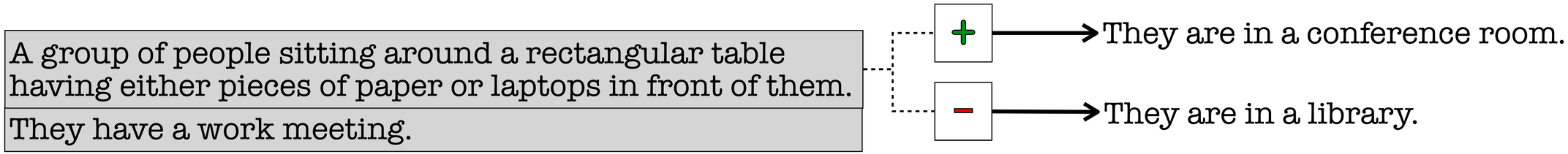
Discriminative Task

Determine whether an update weakens or strengthens the hypothesis.



Generative Task

Generate a weakening or strengthening update for a given premise-hypothesis pair.

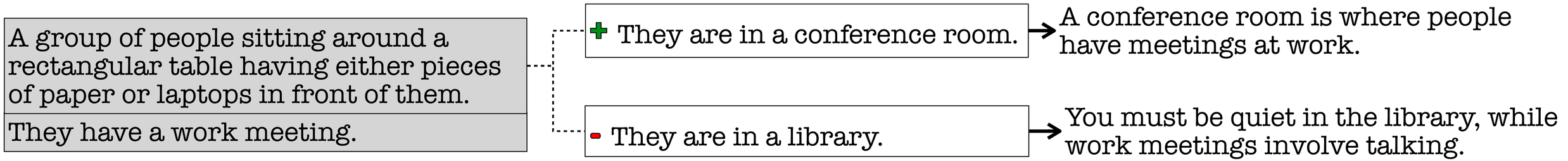


Language models leave plenty of room for improvement on the generative task!

Rationale Generation for Defeasible Inference

Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision. Faeze Brahman, **Vered Shwartz**, Rachel Rudinger, and Yejin Choi. AAAI 2021.

Rationale Generation for Defeasible Inference



Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision. Faeze Brahman, **Vered Shwartz**, Rachel Rudinger, and Yejin Choi. AAAI 2021.

Rationale Generation for Defeasible Inference

A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting.

+ They are in a conference room. → A conference room is where people have meetings at work.

- They are in a library. → You must be quiet in the library, while work meetings involve talking.

Distant supervision:

Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision. Faeze Brahman, **Vered Shwartz**, Rachel Rudinger, and Yejin Choi. AAAI 2021.

Rationale Generation for Defeasible Inference

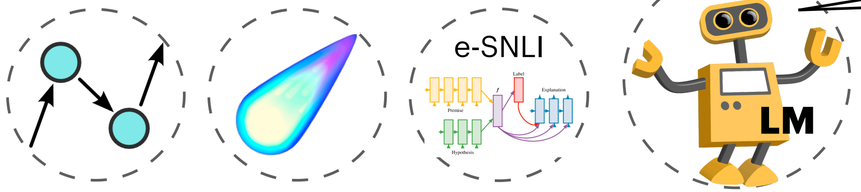
A group of people sitting around a rectangular table having either pieces of paper or laptops in front of them.

They have a work meeting.

+ They are in a conference room. → A conference room is where people have meetings at work.

- They are in a library. → You must be quiet in the library, while work meetings involve talking.

Distant supervision:

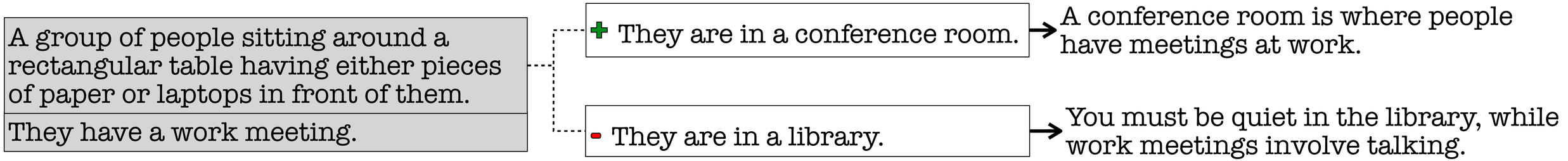


Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision. Faeze Brahman, **Vered Shwartz**, Rachel Rudinger, and Yejin Choi. AAAI 2021.

Rationale Generation for Defeasible Inference

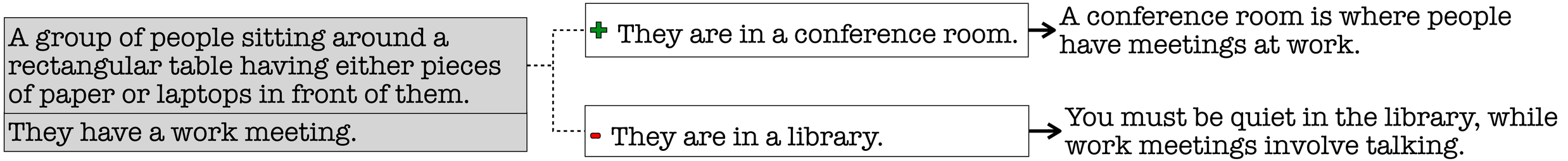
Rationale Generation for Defeasible Inference

Post hoc Rationalization **Generates a rationale for a given decision (label).**



Rationale Generation for Defeasible Inference

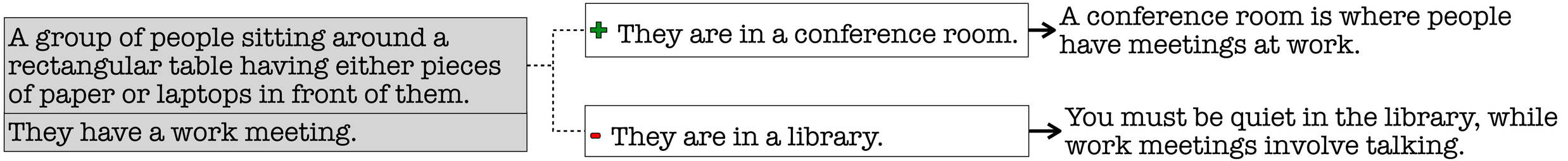
Post hoc Rationalization **Generates a rationale for a given decision (label).**



Trivially rephrasing the label! (“[+] implies that [H]”)

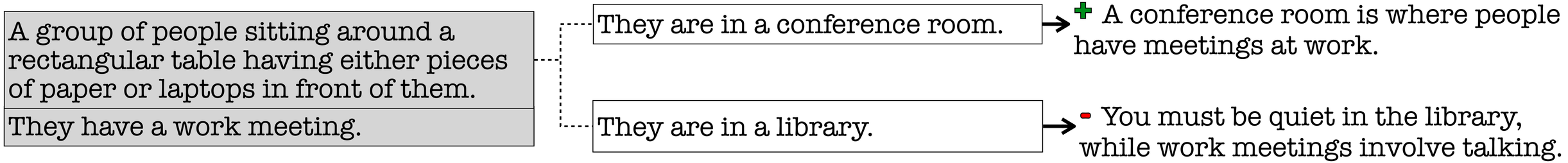
Rationale Generation for Defeasible Inference

Post hoc Rationalization **Generates a rationale for a given decision (label).**



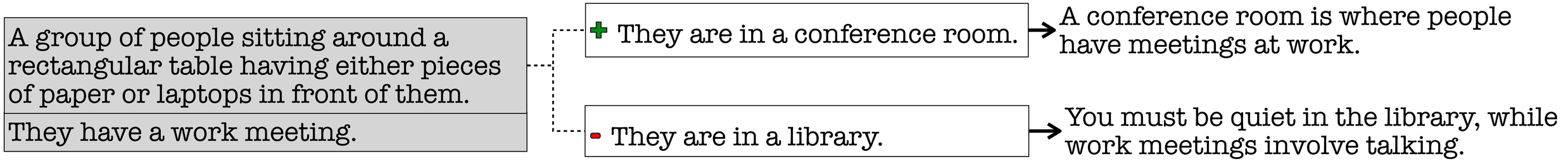
Trivially rephrasing the label! ("[+] implies that [H]")

Joint Prediction & Rationalization **Predict the label (strengthener / weakener) and rationalize it.**



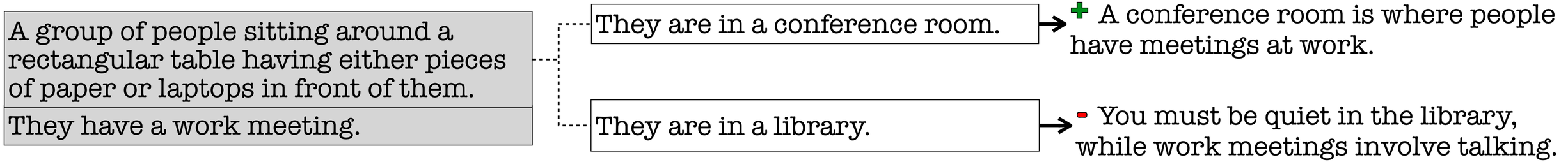
Rationale Generation for Defeasible Inference

Post hoc Rationalization **Generates a rationale for a given decision (label).**



Trivially rephrasing the label! (“[+] implies that [H]”)

Joint Prediction & Rationalization **Predict the label (strengthener / weakener) and rationalize it.**



More realistic but very challenging task!

Recap

🤖 Neural models achieve impressive gains on NLI
- But make stupid unhuman like errors
- “Human performance” is debatable

Recap

- Neural models achieve impressive gains on NLI
  - But make stupid unhuman like errors
  - “Human performance” is debatable

- Symbolic knowledge is useful

Recap

- Neural models achieve impressive gains on NLI
  - But make stupid unhuman like errors
  - “Human performance” is debatable

- Symbolic knowledge is useful
  - Accurate

Recap

🤖 Neural models achieve impressive gains on NLI
- But make stupid unhuman like errors
- “Human performance” is debatable

🤖 Symbolic knowledge is useful
- Accurate
- More efficient than “learning all the facts”

Recap

🤖 Neural models achieve impressive gains on NLI
- But make stupid unhuman like errors
- “Human performance” is debatable

🤖 Symbolic knowledge is useful
- Accurate
- More efficient than “learning all the facts”
- Combination is not trivial

Recap

- Neural models achieve impressive gains on NLI
  - But make stupid unhuman like errors
  - “Human performance” is debatable
- Symbolic knowledge is useful
  - Accurate
  - More efficient than “learning all the facts”
  - Combination is not trivial
- Time to work on more real-world NLI tasks
  - Partial and defeasible inferences

Recap

🤖 Neural models achieve impressive gains on NLI
- But make stupid unhuman like errors
- "Human performance" is debatable

🤖 Symbolic knowledge is useful
- Accurate
- More efficient than "learning all the facts"
- Combination is not trivial

🤖 Time to work on more real-world NLI tasks
- Partial and defeasible inferences

Thank You!

@VeredShwartz vereds@allenai.org

References (1)

(1) Vered Shwartz and Yejin Choi. Do Neural Language Models Overcome Reporting Bias? COLING 2020.

(2) Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, Benjamin Van Durme. Hypothesis Only Baselines in Natural Language Inference. *SEM 2018.

(3) Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. TACL 2020

(4) Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander Miller. Language Models as Knowledge Bases? EMNLP 2019.

(5) Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. Workshop on Automated knowledge base construction 2013.

(6) Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. ACL 2020.

(7) Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, Noah A. Smith. Annotation Artifacts in Natural Language Inference Data. NAACL 2018

(8) Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. AAAI 2020.

(9) Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. Probing Neural Language Models for Human Tacit Assumptions. CogSci 2020.

(10) Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How Can We Know What Language Models Know? TACL 2020.

(11) Maxwell Forbes, Ari Holtzman, and Yejin Choi. Do Neural Language Representations Learn Physical Commonsense? CogSci 2019.

(12) Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, Colin Raffel. Extracting Training Data from Large Language Models. arXiv 2020.

(13) Masatoshi Tsuchiya. Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment. In 11th International Conference on Language Resources and Evaluation (LREC2018), 2018.

(14) R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL 2019.

(15) Kyle Richardson, Hai Hu, Lawrence S Moss, Ashish Sabharwal. Probing Natural Language Inference Models through Semantic Fragments. AAAI 2020.

References (2)

(16) Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. TACL 2019.

(17) Raymond Reiter. A Logic for Default Reasoning. Artificial Intelligence, 1980.

(18) Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. Neurips 2019.

(19) Max Glockner, Vered Shwartz, Yoav Goldberg. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. ACL 2018.

(20) Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, Graham Neubig. Stress Test Evaluation for Natural Language Inference. COLING 2018.

(21) Ido Dagan, Dan Roth, Mark Sammons, and Fabio Zanzotto. Recognizing Textual Entailment. Morgan & Claypool Publishers, 2013.

(22) Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. AAAI 2020.

(23) Zhengbao Jiang, Frank F. Xu, Jun Araki, Graham Neubig. How Can We Know What Language Models Know? TACL 2019.

(24) Nelson F. Liu, Roy Schwartz, and Noah A. Smith. Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. NAACL 2019.

(25) Ohad Rozen, Vered Shwartz, Roee Aharoni, and Ido Dagan. Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets. CoNLL 2019.

(26) Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge Enhanced Contextual Word Representations. EMNLP 2019.

(27) Ohad Rozen, Shmuel Amar, Vered Shwartz and Ido Dagan. Teach the Rules, Provide the Facts: Targeted Relational-knowledge Enhancement for Textual Inference. *SEM 2021.

(28) Rachel Rudinger, Vered Shwartz, Jena Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah Smith, and Yejin Choi. Thinking Like a Skeptic: Defeasible Inference in Natural Language. Findings of EMNLP 2020.

(29) Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. Learning to Rationalize for Nonmonotonic Reasoning with Distant Supervision. AAAI 2021.