Fast Breaking and Slow Building
of textual inference models

Vered Shwartz

December 2019

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

State-of-the-art AI solutions: (1) Google BERT, an AI model that understands language better than humans

AIN Dev Team Follow
Jan 31 · 8 min read

VICE

Algorithms Have Nearly Mastered Human Language. Why Can't They Stop Being Sexist?

It turns out that data-fueled algorithms are no better than humans—and ... Even AI researchers who work with machine learning models—like neural nets, which ...

Sep 18, 2019

Alizila

Alibaba AI Beats Humans in Reading-Comprehension...

Alibaba Group's machine-learning technology is better at reading comprehension than humans, according to a well-known test built for the industry by Microsoft.

Jul 9, 2019


shutterstock.com • 1022694991

What's in this talk?

	Breaking	Building
NLI	[Glockner et al., 2018] [Rozen et al., 2019]	
Coreference		[Shwartz et al., 2017] [Barhom et al., 2019]

	Breaking	Building
NLI	[Glockner et al., 2018] [Rozen et al., 2019]	
Coreference		

Breaking NLI Systems
with Sentences that Require Simple Lexical Inferences

Max Glockner[1], **Vered Shwartz**[2] **and Yoav Goldberg**[2]

[1]TU Darmstadt

[2]Bar-Ilan University

TECHNISCHE
UNIVERSITÄT
DARMSTADT

B I U
N L P

ACL 2018

SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])

SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

Premise

Street performer is doing his act for kids

Hypotheses

SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

Premise

Street performer is doing his act for kids

Hypotheses

1. A person performing for children on the street

SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

Premise

Street performer is doing his act for kids

Hypotheses

1. A person performing for children on the street) ENTAILMENT

SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

Premise

Street performer is doing his act for kids

Hypotheses

1. A person performing for children on the street) ENTAILMENT
2. A juggler entertaining a group of children on the street

SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

Premise

Street performer is doing his act for kids

Hypotheses

1. A person performing for children on the street) ENTAILMENT
2. A juggler entertaining a group of children on the street) NEUTRAL

SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

Premise

Street performer is doing his act for kids

Hypotheses

1. A person performing for children on the street) ENTAILMENT
2. A juggler entertaining a group of children on the street) NEUTRAL
3. A magician performing for an audience in a nightclub

SNLI [Bowman et al., 2015]

- A large scale dataset for NLI (Natural Language Inference; Recognizing Textual Entailment [Dagan et al., 2013])
- Premises are image captions, hypotheses generated by crowdsourcing workers:

Premise

Street performer is doing his act for kids

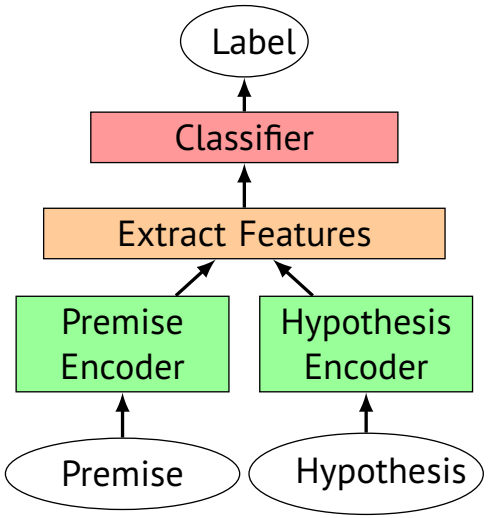
Hypotheses

1. A person performing for children on the street) ENTAILMENT
2. A juggler entertaining a group of children on the street) NEUTRAL
3. A magician performing for an audience in a nightclub) CONTRADICTION

- Event co-reference assumption

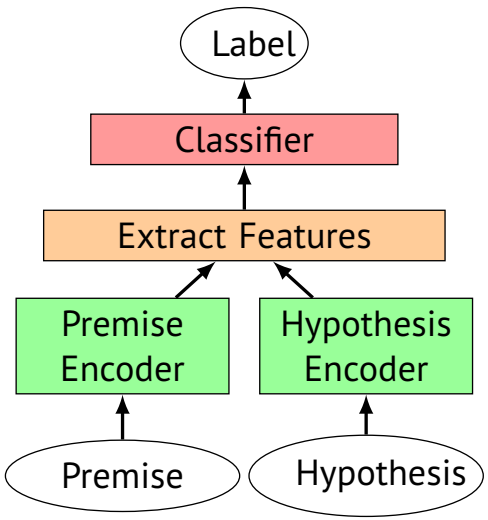
Neural NLI Models

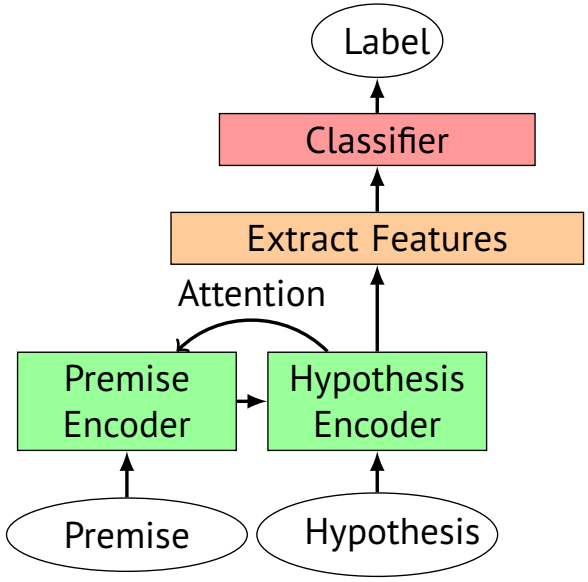
- End-to-end, either **sentence-encoding** or **attention-based**



Neural NLI Models

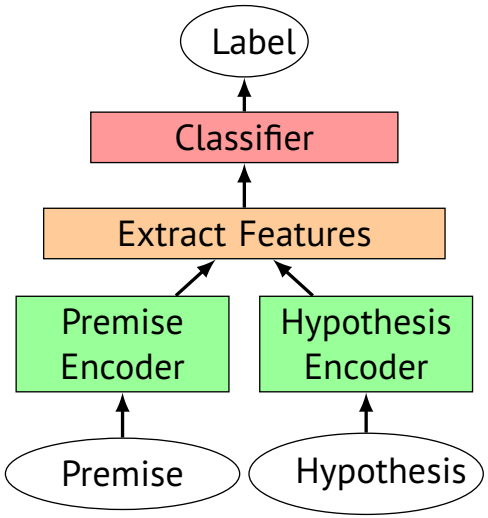
- End-to-end, either **sentence-encoding** or **attention-based**

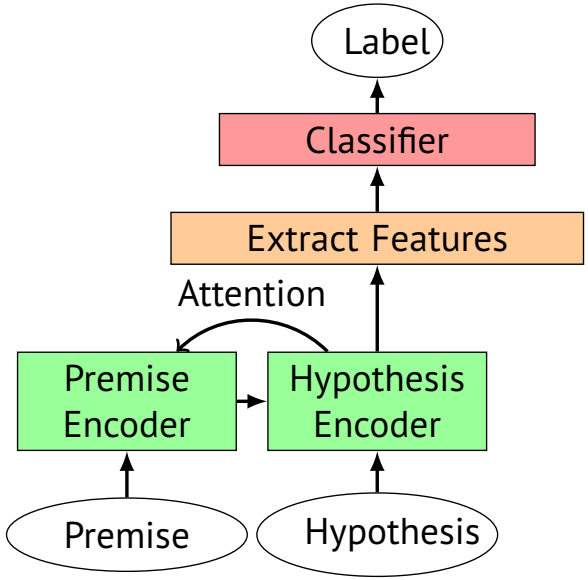




Neural NLI Models

- End-to-end, either **sentence-encoding** or **attention-based**

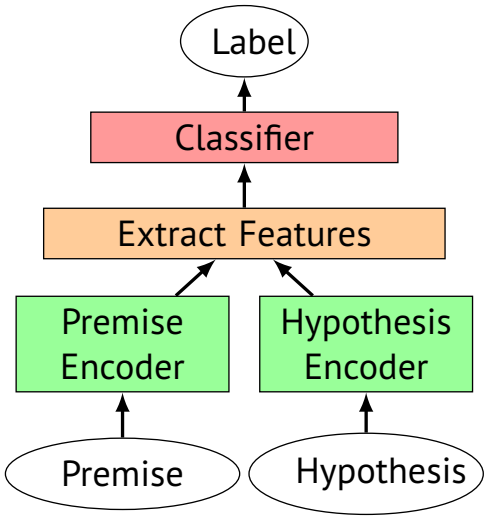


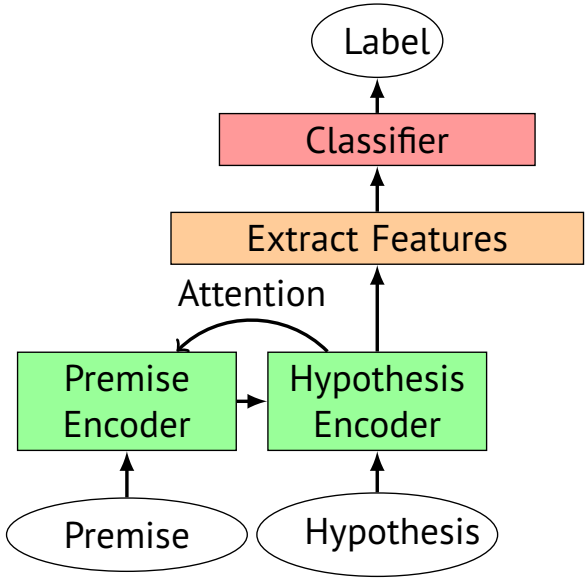


- Lexical knowledge: only from pre-trained word embeddings
  - As opposed to using resources like WordNet

Neural NLI Models

- End-to-end, either **sentence-encoding** or **attention-based**

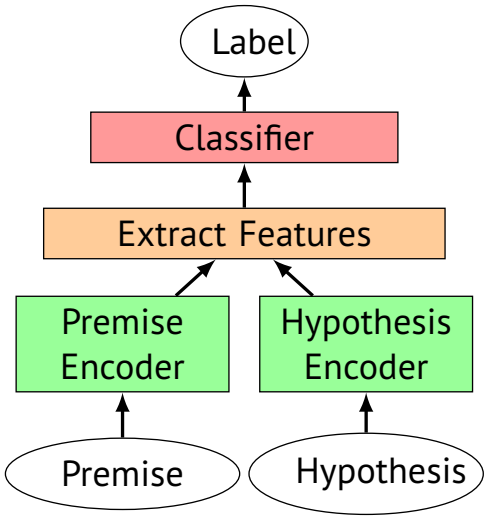


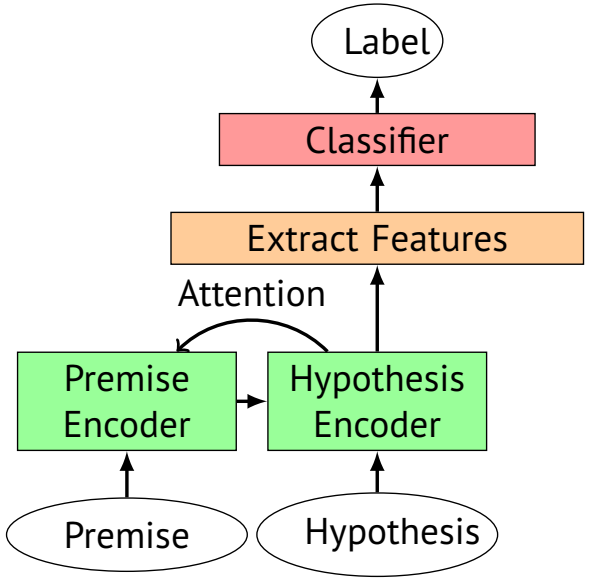


- Lexical knowledge: only from pre-trained word embeddings
  - As opposed to using resources like WordNet
- SOTA exceeds human performance…

Neural NLI Models

- End-to-end, either **sentence-encoding** or **attention-based**



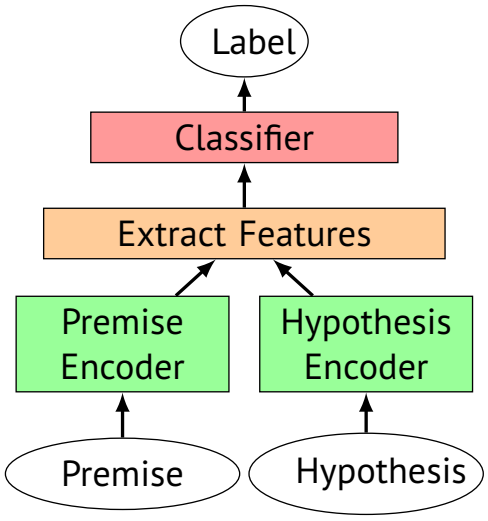


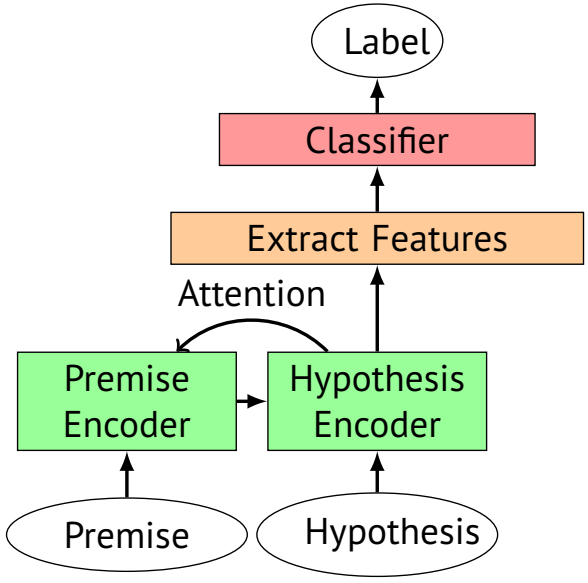
- Lexical knowledge: only from pre-trained word embeddings
  - As opposed to using resources like WordNet
- SOTA exceeds human performance... [1]

[1]

Neural NLI Models

- End-to-end, either **sentence-encoding** or **attention-based**





- Lexical knowledge: only from pre-trained word embeddings
  - As opposed to using resources like WordNet
- SOTA exceeds human performance... [1]

[1] [Gururangan et al., 2018, Poliak et al., 2018]: by learning "easy clues"

Annotation Artifacts in SNLI

- [Gururangan et al., 2018, Poliak et al., 2018]: good performance on SNLI based on the hypothesis alone

Annotation Artifacts in SNLI

- [Gururangan et al., 2018, Poliak et al., 2018]: good performance on SNLI based on the hypothesis alone
- This is a result of the annotation procedure

Annotation Artifacts in SNLI

- [Gururangan et al., 2018, Poliak et al., 2018]: good performance on SNLI based on the hypothesis alone
- This is a result of the annotation procedure
  - Negation (not, never, nobody) is correlated with *contradiction*

Annotation Artifacts in SNLI

- [Gururangan et al., 2018, Poliak et al., 2018]: good performance on SNLI based on the hypothesis alone
- This is a result of the annotation procedure
  - Negation (not, never, nobody) is correlated with *contradiction*
  - …and “cat” as well (many dog images)

Annotation Artifacts in SNLI

- [Gururangan et al., 2018, Poliak et al., 2018]: good performance on SNLI based on the hypothesis alone
- This is a result of the annotation procedure
  - Negation (not, never, nobody) is correlated with *contradiction*
  - …and “cat” as well (many dog images)
  - Generic words (animal, instrument) are correlated with *entailment*

Annotation Artifacts in SNLI

- [Gururangan et al., 2018, Poliak et al., 2018]: good performance on SNLI based on the hypothesis alone
- This is a result of the annotation procedure
  - Negation (not, never, nobody) is correlated with *contradiction*
  - …and "cat" as well (many dog images)
  - Generic words (animal, instrument) are correlated with *entailment*
  - Sentence length: *entailment* < *contradiction* < *neutral*

Do neural NLI models implicitly learn lexical semantic relations?

New Test Set

- We constructed a new test set to answer this question

New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set

New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set
- **Hypothesis**:
  - Replacing a single term w in the premise with a related term w^0

New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set
- **Hypothesis**:
  - Replacing a single term w in the premise with a related term w'
  - w' is in the SNLI vocabulary and in pre-trained embeddings

New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set
- **Hypothesis**:
  - Replacing a single term w in the premise with a related term w'
  - w' is in the SNLI vocabulary and in pre-trained embeddings
  - Crowdsourcing labels (mostly contradictions!)

New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set
- **Hypothesis**:
  - Replacing a single term w in the premise with a related term w^0
  - w^0 is in the SNLI vocabulary and in pre-trained embeddings
  - Crowdsourcing labels (mostly contradictions!)

Contradiction

The man is holding a saxophone *!* The man is holding an electric guitar

New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set
- **Hypothesis**:
  - Replacing a single term w in the premise with a related term w'
  - w' is in the SNLI vocabulary and in pre-trained embeddings
  - Crowdsourcing labels (mostly contradictions!)

Contradiction

The man is holding a saxophone *!* The man is holding an electric guitar

Entailment

A little girl is very sad *!* A little girl is very unhappy

New Test Set

- We constructed a new test set to answer this question
- **Premise**: sentences from the SNLI training set
- **Hypothesis**:
  - Replacing a single term w in the premise with a related term w^0
  - w^0 is in the SNLI vocabulary and in pre-trained embeddings
  - Crowdsourcing labels (mostly contradictions!)

Contradiction

The man is holding a saxophone *!* The man is holding an electric guitar

Entailment

A little girl is very sad *!* A little girl is very unhappy

Neutral

A couple drinking wine *!* A couple drinking champagne

Evaluation Setting

- 3 representative models:
  - `Residual-Stacked-Encoder` [Nie and Bansal, 2017]
  - `ESIM` (Enhanced Sequential Inference Model) [Chen et al., 2017]
  - `Decomposable Attention` [Parikh et al., 2016]

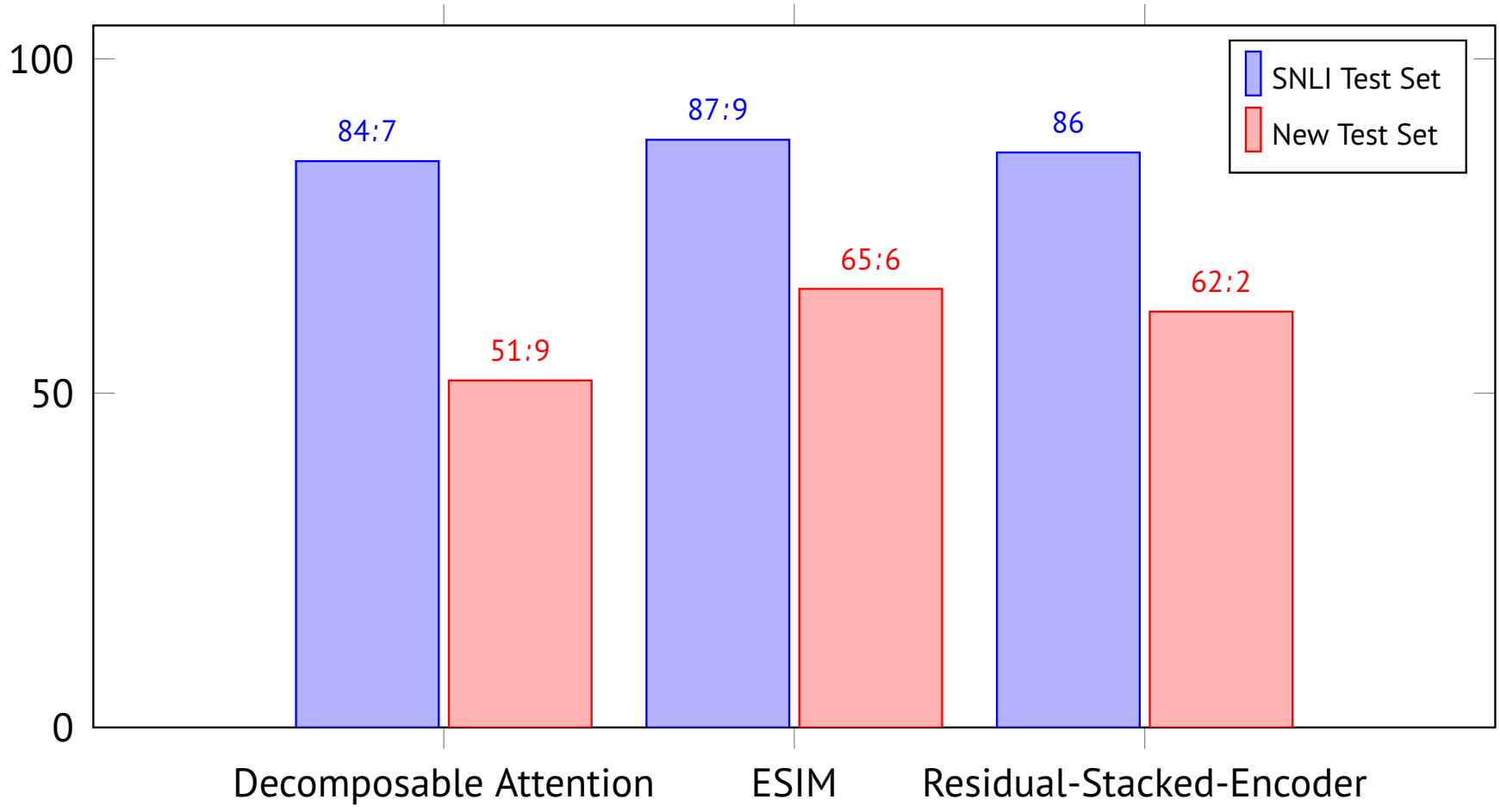
Evaluation Setting

- 3 representative models:
  - `Residual-Stacked-Encoder` [Nie and Bansal, 2017]
  - `ESIM` (Enhanced Sequential Inference Model) [Chen et al., 2017]
  - `Decomposable Attention` [Parikh et al., 2016]

- Train on SNLI training set, test on the original & new test set
  - In the paper: enhancing with additional existing datasets

Results

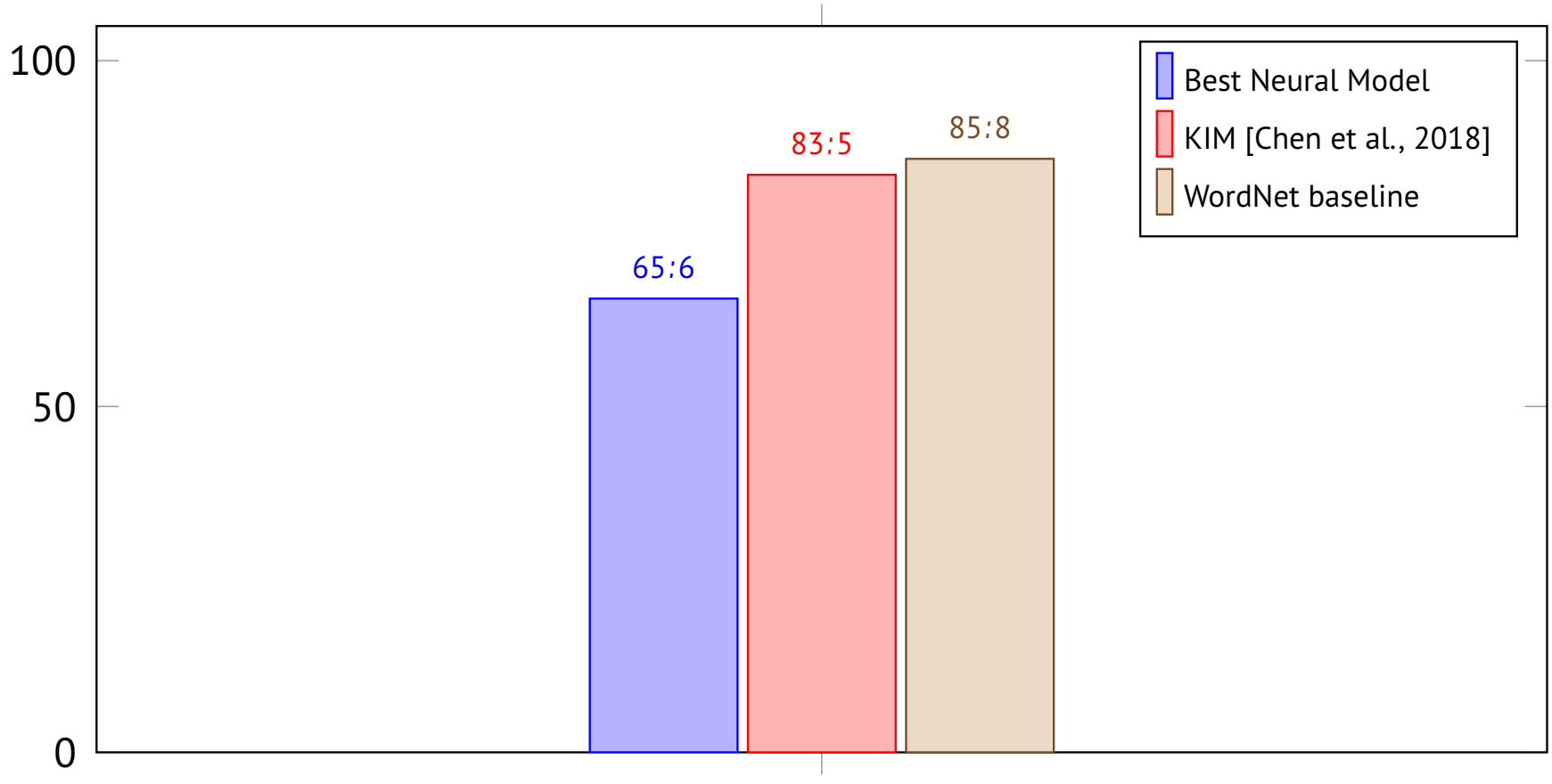
Can neural NLI models recognize lexical inferences?



Dramatic drop in performance across models.

Sanity Check

Performance of WordNet-informed Models



The test set is solvable using WordNet.

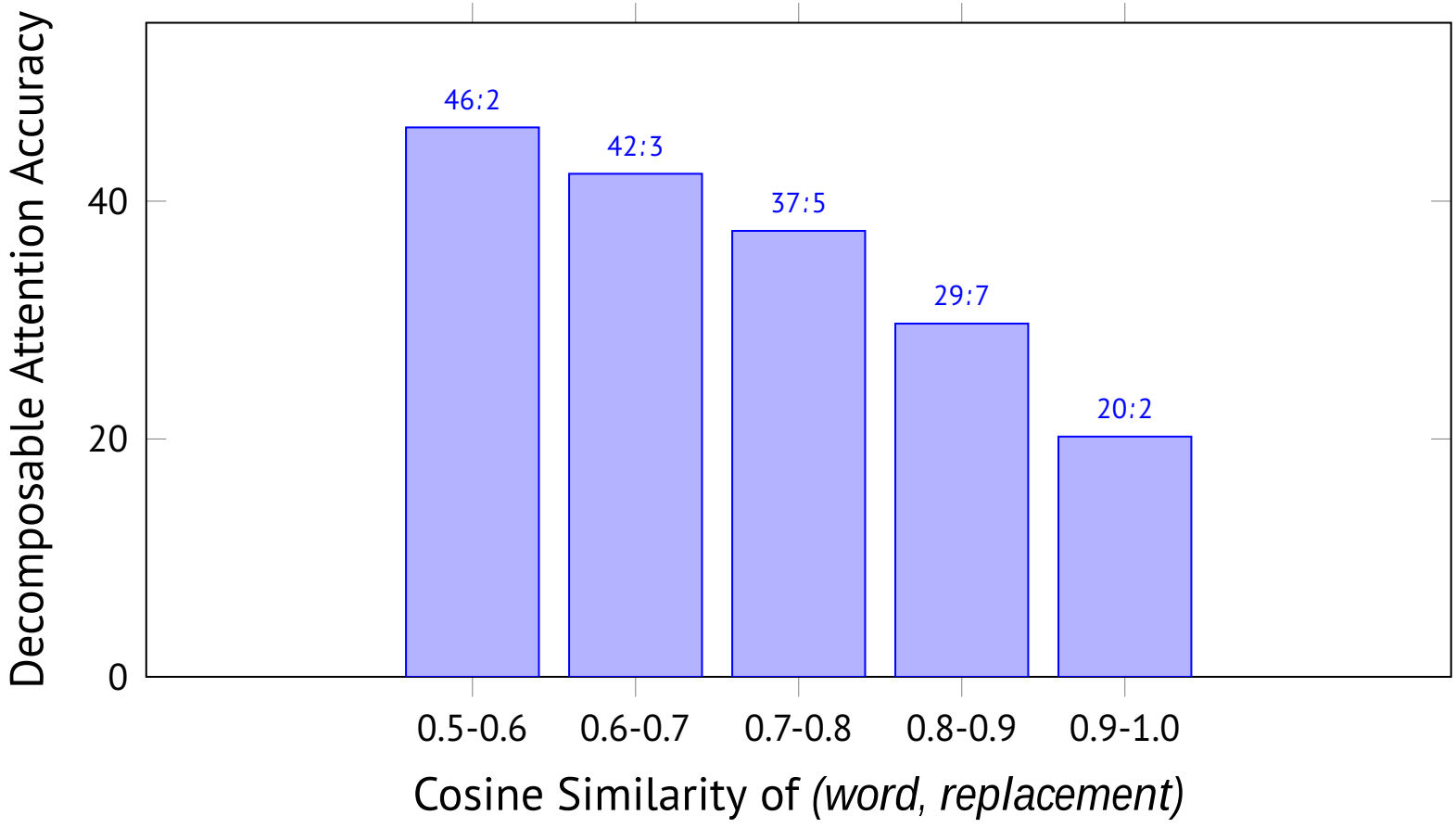
What do neural NLI models learn with respect to lexical semantic relations?

Analysis 1: Word Similarity

- Models err on contradicting word-pairs with similar embeddings
  - *A man starts his day in India ! A man starts his day in Malaysia*

Analysis 1: Word Similarity

- Models err on contradicting word-pairs with similar embeddings
  - *A man starts his day in India ! A man starts his day in Malaysia*
- Especially for fixed word embeddings

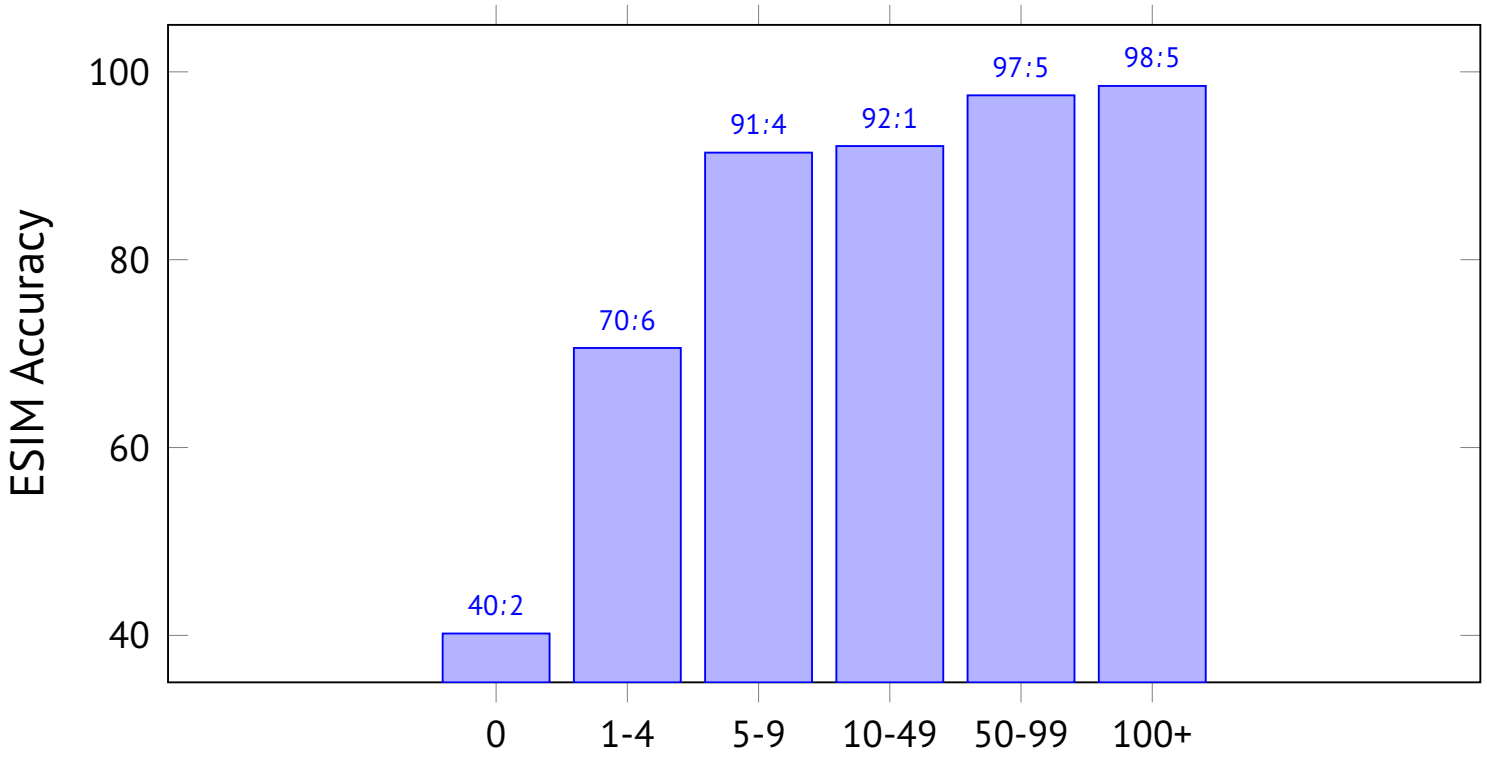


Analysis 2: Frequency in Training

- Tuning embeddings may associate specific *(word, replacement)* pairs to a label, e.g. *(man, woman)* ! contradiction

Analysis 2: Frequency in Training

- Tuning embeddings may associate specific *(word, replacement)* pairs to a label, e.g. *(man, woman)* ! contradiction
- Accuracy increases with frequency in training set



Breaking NLI

Recap

- New NLI test set that evaluates systems' ability to make inferences that require *very simple* lexical knowledge

Breaking NLI

Recap

- New NLI test set that evaluates systems' ability to make inferences that require *very simple* lexical knowledge

- SOTA systems perform poorly on the test set

Breaking NLI

Recap

- New NLI test set that evaluates systems' ability to make inferences that require *very simple* lexical knowledge
- SOTA systems perform poorly on the test set
- Systems are limited in their generalization ability

Breaking NLI

Recap

- New NLI test set that evaluates systems' ability to make inferences that require *very simple* lexical knowledge
- SOTA systems perform poorly on the test set
- Systems are limited in their generalization ability
- **Related Work**:
  - "Stress tests" [Naik et al., 2018]: similar findings on a broader range of linguistic phenomena

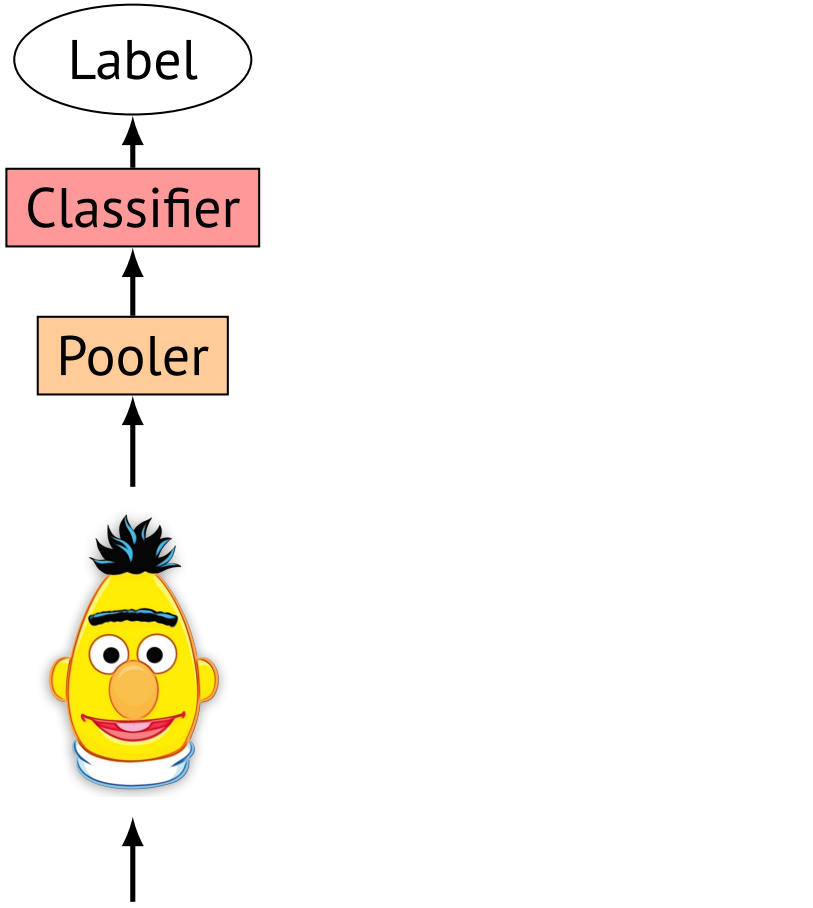
Breaking NLI

Recap

- New NLI test set that evaluates systems' ability to make inferences that require *very simple* lexical knowledge
- SOTA systems perform poorly on the test set
- Systems are limited in their generalization ability
- **Related Work**:
  - "Stress tests" [Naik et al., 2018]: similar findings on a broader range of linguistic phenomena
  - Inference with single word differences:
    [Pavlick and Callison-Burch, 2016, Kalouli et al., 2018]

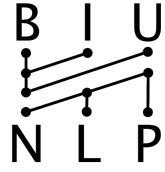
But current LM-based NLI models address entailment-related phenomena better, no?

Pre-trained LM based NLI models



Diversify Your Datasets:
Analyzing Generalization via Controlled Variance in Adversarial Datasets

Ohad Rozen, Vered Shwartz, Roee Aharoni, and Ido Dagan



Bar-Ilan University

CoNLL 2019

MultiNLI [Williams et al., 2018]

- Collected like SNLI (existing premises, generated hypotheses)

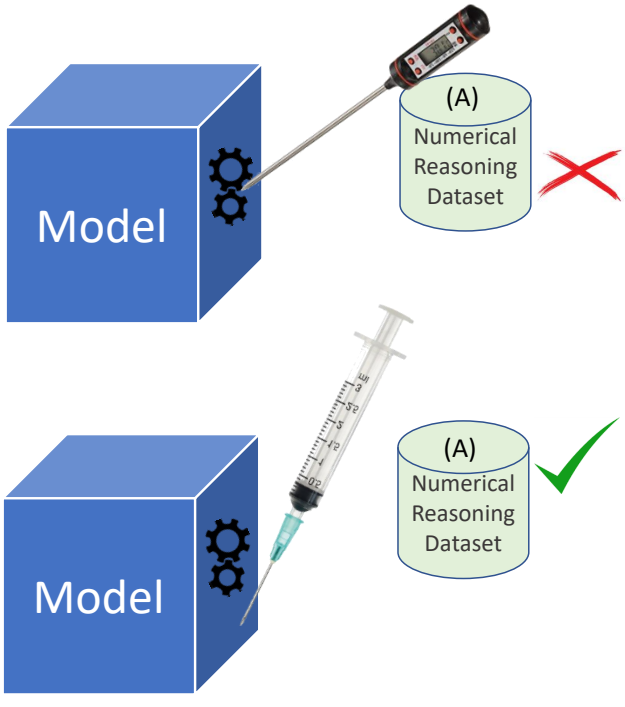
MultiNLI [Williams et al., 2018]

- Collected like SNLI (existing premises, generated hypotheses)
- Multiple geners

MultiNLI [Williams et al., 2018]

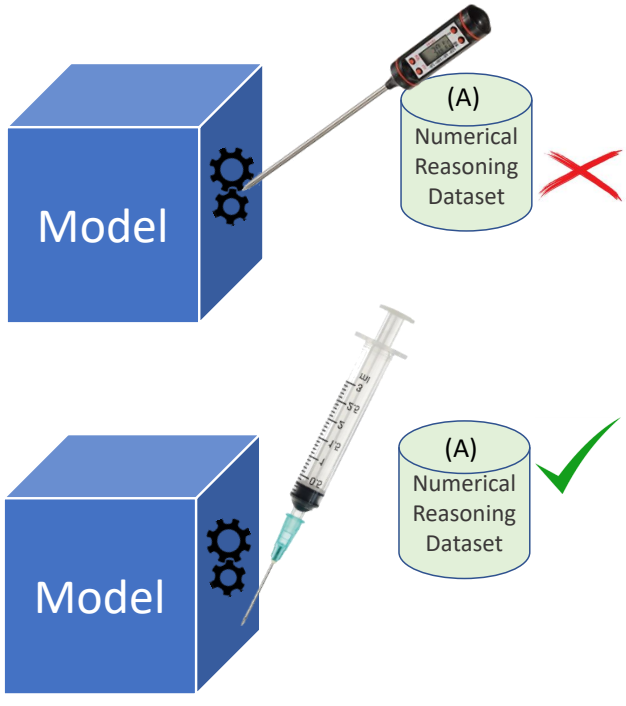
- Collected like SNLI (existing premises, generated hypotheses)
- Multiple geners
- Mismatched evaluation (not in our focus)

Probing ! Inoculation



- **Probing**: does the representation *capture* a certain property? [Glockner et al., 2018, Naik et al., 2018]

Probing ! Inoculation



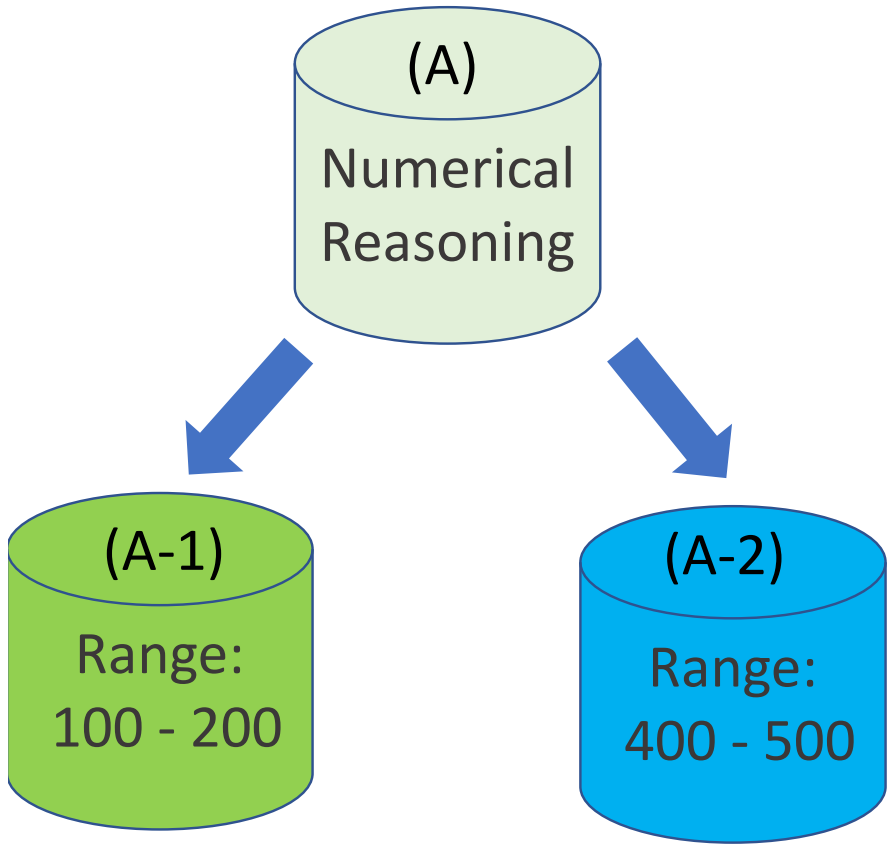
- **Probing**: does the representation *capture* a certain property? [Glockner et al., 2018, Naik et al., 2018]
- **Inoculation** [Liu et al., 2019]: can the representation *learn* a certain property?

Has the model learned a general notion of the property, or does it overfit to the specific dataset?

Testing Generalization Capacity

Methodology

1. Split the challenge dataset to different variations across the dimension in focus.



Testing Generalization Capacity

Methodology

2. Fine-tune on one set and test on another.

Testing Generalization Capacity

Methodology

Testing Generalization Capacity

Methodology

Challenge Datasets Generation

Templates from MultiNLI sentences!
Ask crowdsourcing workers to rephrase spans.

Phenomena

Dative Alternation

Phenomena

Numeric Reasoning

Challenge Datasets Generation

Diversity Axes

1. Syntax complexity - simple / medium / complex
   - Sentence length
   - Phenomenon depth in parse tree

Challenge Datasets Generation

Diversity Axes

1. Syntax complexity - simple / medium / complex
   - Sentence length
   - Phenomenon depth in parse tree
2. Lexical variability
   - Dative verb
   - Number range

Inoculation tells part of the story...

...our data tells the other part

Dative Alternation

Not sensitive to lexical variability

...our data tells the other part

Dative Alternation

Not sensitive to lexical variability

But generalizes only from complex to simple syntax

...our data tells the other part

Dative Alternation

Not sensitive to lexical variability

But generalizes only from complex to simple syntax

Numeric Reasoning

Not sensitive to syntax

...our data tells the other part

Dative Alternation
Not sensitive to lexical variability
But generalizes only from complex to simple syntax

Numeric Reasoning
Not sensitive to syntax
But doesn't generalize across number ranges

Diversify your Datasets

Recap

- Simple methodology to test model generalization of a speci c learned phenomenon

Diversify your Datasets

Recap

- Simple methodology to test model generalization of a speci c learned phenomenon
- NLI-BERT fails to generalize dative alternation and numeric reasoning

Diversify your Datasets

Recap

- Simple methodology to test model generalization of a speci c learned phenomenon
- NLI-BERT fails to generalize dative alternation and numeric reasoning
- Fine-tuning on the phenomenon-speci c data may decrease the main task performance (also in [Richardson et al., 2020]).

Real-world examples: partial entailments

S_1: Amazon To Acquire Whole Foods Marke For $13.7 Billion

S_2: Amazon is buying Whole Foods for almost $14 billion in cash

Real-world examples: partial entailments

S_1: Researchers have discovered wreckage of the lost warship , the USS Indianapolis after 72 years

S_2: Wreckage of missing WWII ship found in Pacific Ocean

	Breaking	Building
NLI		
Coreference		[Shwartz et al., 2017] [Barhom et al., 2019]

Revisiting Joint Modeling of Cross-document Entity and Event Coreference Resolution

Shany Barhom[1], Vered Shwartz[1], Alon Eirew[2],
Michael Bugert[3], Nils Reimers[3], and Ido Dagan[1]

[1]Bar-Ilan University [2]Intel AI Lab [3]TU Darmstadt

ACL 2019

Joint Event and Entity Coreference Resolution

Recap

- Cross-document coreference is drastically under-explored

Joint Event and Entity Coreference Resolution

Recap

- Cross-document coreference is drastically under-explored
- A simple joint approach with state-of-the-art results on ECB+

Joint Event and Entity Coreference Resolution

Recap

- Cross-document coreference is drastically under-explored
- A simple joint approach with state-of-the-art results on ECB+
- Still a long way to go!

Acquiring Predicate Paraphrases from News Tweets

Vered Shwartz, Gabriel Stanovsky, and Ido Dagan

Bar-Ilan University

*SEM 2017

Acquiring Predicate Paraphrases from News Tweets[2]

[a]$_0$ introduce [a]$_1$	[a]$_0$ welcome [a]$_1$
[a]$_0$ appoint [a]$_1$	[a]$_0$ to become [a]$_1$
[a]$_0$ die at [a]$_1$	[a]$_0$ pass away at [a]$_1$
[a]$_0$ hit [a]$_1$	[a]$_0$ sink to [a]$_1$
[a]$_0$ be investigate [a]$_1$	[a]$_0$ be probe [a]$_1$
[a]$_0$ eliminate [a]$_1$	[a]$_0$ slash [a]$_1$
[a]$_0$ announce [a]$_1$	[a]$_0$ unveil [a]$_1$
[a]$_0$ quit after [a]$_1$	[a]$_0$ resign after [a]$_1$
[a]$_0$ announce as [a]$_1$	[a]$_0$ to become [a]$_1$
[a]$_0$ threaten [a]$_1$	[a]$_0$ warn [a]$_1$
[a]$_0$ die at [a]$_1$	[a]$_0$ live until [a]$_1$
[a]$_0$ double down on [a]$_1$	[a]$_0$ stand by [a]$_1$
[a]$_0$ kill [a]$_1$	[a]$_0$ shoot [a]$_1$
[a]$_0$ approve [a]$_1$	[a]$_0$ pass [a]$_1$
seize [a]$_0$ at [a]$_1$	to grab [a]$_0$ at [a]$_1$

- Binary verbal predicate paraphrases
- Extracted from Twitter
- Ever-growing resource: currently around 5.2M paraphrases

[2] Available at https://github.com/vered1986/Chirps

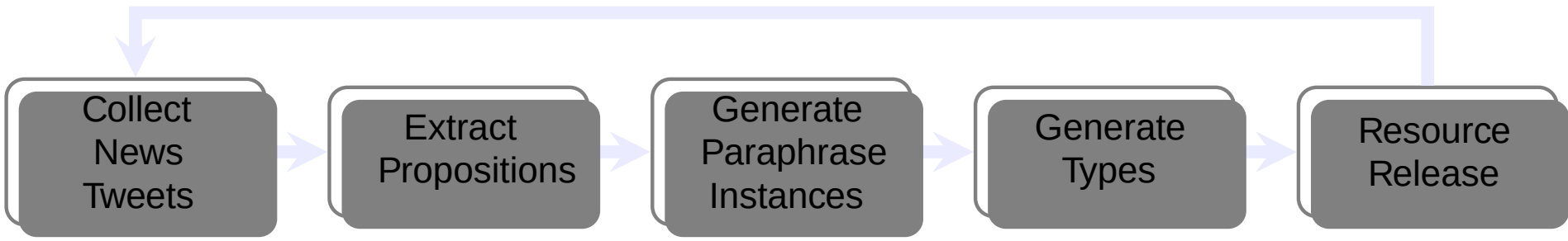
Assumptions

- Main assumption:redundant news headlines of the same event are likely to describe it with different words [Shinyama et al., 2002, Barzilay and Lee, 2003].

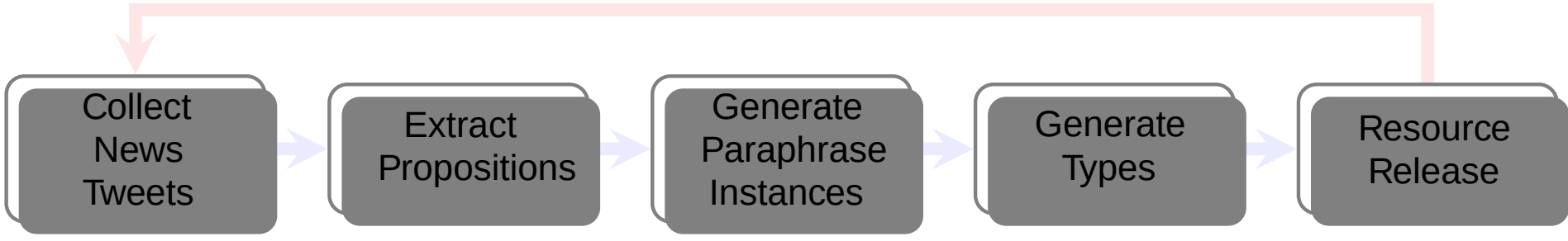
Assumptions

- Main assumption:redundant news headlines of the same event are likely to describe it with different words [Shinyama et al., 2002, Barzilay and Lee, 2003].
- This work: propositions extracted from tweets discussing news events, published on the same day, that agree on their arguments, are predicate paraphrases.

Resource Collection



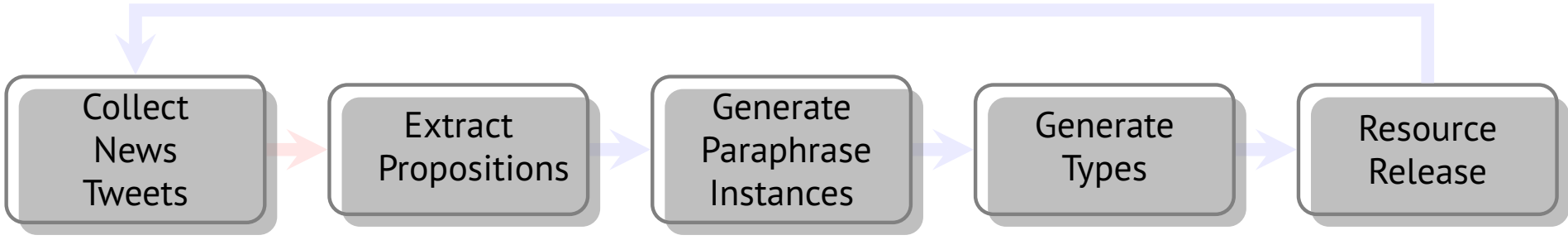
Resource Collection



- Query the Twitter Search API for news tweets in English

Amazon is buying Whole Foods in $13.7B
Amazon to acquire Whole Foods Market in deal valued at nearly $14 billion

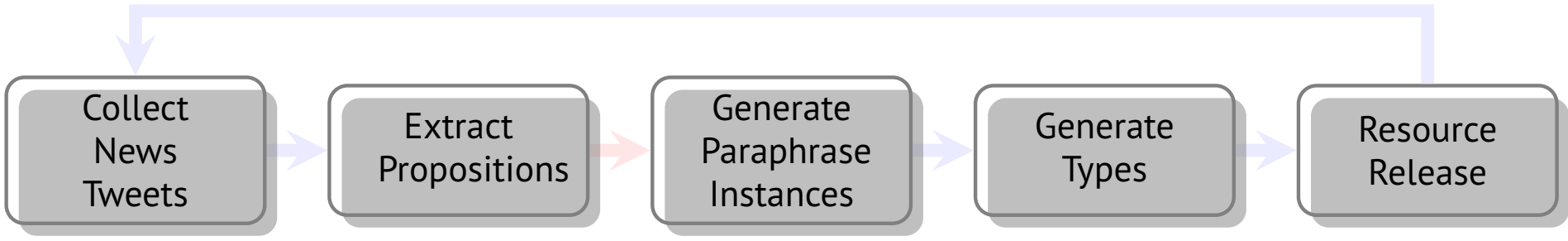
Resource Collection



- Extract propositions from tweets using PropS [Stanovsky et al., 2016]
- Get binary verbal predicate templates, and apply argument reduction [Stanovsky and Dagan, 2016]

[Amazon] **buy** [Whole Foods]
[Amazon] **acquire** [Whole Foods Market]

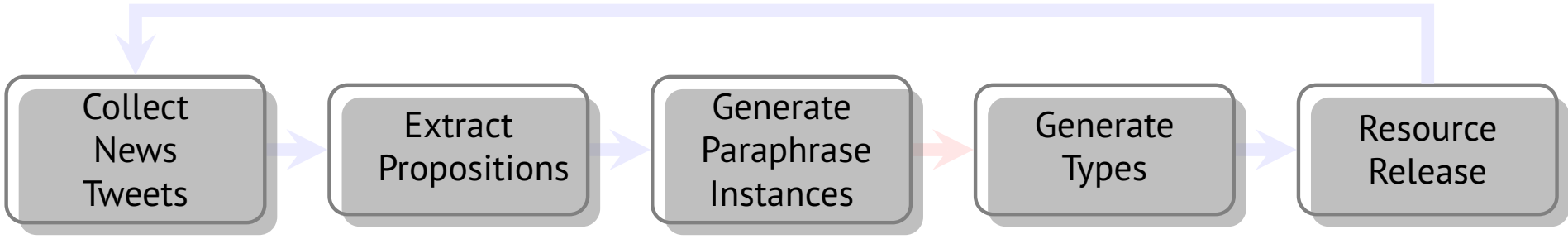
Resource Collection



- We consider two predicates as paraphrases if:
  1. They appear on the same day.
  2. Each of their arguments aligns with a unique argument in the other predicate.
- Two levels of argument matching: **strict** (exact match / short edit distance) and **loose** (partial token matching / WordNet synonyms)

$[a]_0$ **buy** $[a]_1$	$[a]_0$ **acquire** $[a]_1$	Amazon	Whole Foods
$[a]_0$ **buy** $[a]_1$	$[a]_0$ **acquire** $[a]_1$	Intel	Mobileye
	...		

Resource Collection



Heuristic score for a predicate paraphrase type:

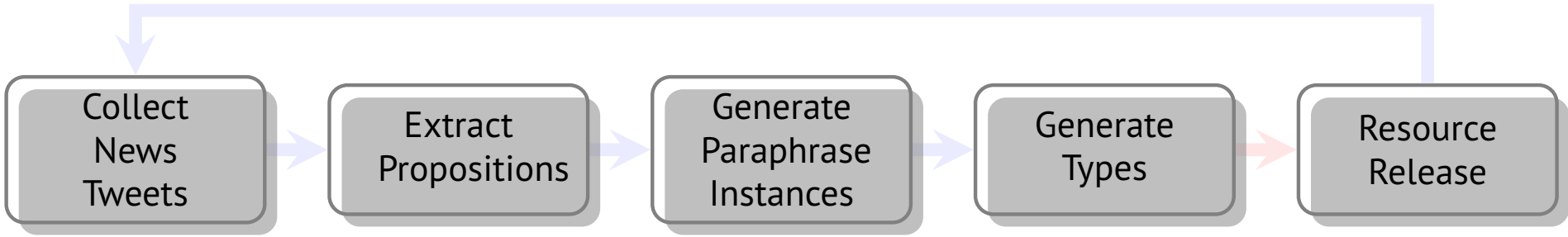
$$p_1 = [a]_0 \textbf{ buy } [a]_1; \quad p_2 = [a]_0 \textbf{ acquire } [a]_1$$

$$s(p_1; p_2) = count(p_1; p_2) \left(1 + \frac{days(p_1, p_2)}{N}\right)$$

- $count(p_1; p_2)$ assigns high scores for frequent paraphrases
- N - number of days since the resource collection begun
- $\frac{days(p_1, p_2)}{N}$ eliminates noise from two arguments participating in different events on the same day

  1) *Last year when Chuck Berry turned 90*; 2) *Chuck Berry dies at 90*

Resource Collection



- We release our resource daily, with two files:
  - **Instances**: predicates, arguments and tweet IDs.
  - **Types**: predicate paraphrase pair types ranked in a descending order according to the heuristic accuracy score.

Chirps
Recap

- Using event coreference to extract paraphrases

Chirps
Recap

- Using event coreference to extract paraphrases
- Complementary to other paraphrasing resources

Chirps
Recap

- Using event coreference to extract paraphrases
- Complementary to other paraphrasing resources
- Useful resource for paraphrasing, event coreference, NLI

Thank you!
Questions?

@VeredShwartz

vereds@allenai.org

References I

[Barhom et al., 2019] Barhom, S., Shwartz, V., Eirew, A., Bugert, M., Reimers, N., and Dagan, I. (2019). Revisiting joint modeling of cross-document entity and event coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189, Florence, Italy. Association for Computational Linguistics.

[Barzilay and Lee, 2003] Barzilay, R. and Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *NAACL*.

[Bowman et al., 2015] Bowman, S. R., Angeli, G., Potts, C., and Manning, D. C. (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. Association for Computational Linguistics.

[Chen et al., 2018] Chen, Q., Zhu, X., Ling, Z.-H., Inkpen, D., and Wei, S. (2018). Neural natural language inference models enhanced with external knowledge. In *The 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

[Chen et al., 2017] Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. (2017). Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

[Dagan et al., 2013] Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. M. (2013). Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

References II

[Glockner et al., 2018] Glockner, M., Shwartz, V., and Goldberg, Y. (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.

[Gururangan et al., 2018] Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, New Orleans, Louisiana.

[Kalouli et al., 2018] Kalouli, A.-L., Real, L., and DePaiva, V. (2018). Wordnet for "easy" textual inferences. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

[Liu et al., 2019] Liu, N. F., Schwartz, R., and Smith, N. A. (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.

[Naik et al., 2018] Naik, A., Ravichander, A., Sadeh, N., Rose, C., and Neubig, G. (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.

References III

[Nie and Bansal, 2017] Nie, Y. and Bansal, M. (2017). Shortcut-stacked sentence encoders for multi-domain inference. *arXiv preprint arXiv:1708.02312*.

[Parikh et al., 2016] Parikh, A., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

[Pavlick and Callison-Burch, 2016] Pavlick, E. and Callison-Burch, C. (2016). Most baies are little and most problems are huge: Compositional entailment in adjective nouns.

[Poliak et al., 2018] Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., and Van Durme, B. (2018). Hypothesis Only Baselines in Natural Language Inference. In *Joint Conference on Lexical and Computational Semantics (StarSem)*.

[Richardson et al., 2020] Richardson, K., Hu, H., Moss, L. S., and Sabharwal, A. (2020). Probing natural language inference models through semantic fragments. In *AAAI*.

[Rozen et al., 2019] Rozen, O., Shwartz, V., Aharoni, R., and Dagan, I. (2019). Diversify your datasets: Analyzing generalization via controlled variance in adversarial datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 196–205, Hong Kong, China. Association for Computational Linguistics.

[Shinyama et al., 2002] Shinyama, Y., Sekine, S., and Sudo, K. (2002). Automatic paraphrase acquisition from news articles. In *HLT*, pages 313–318. Morgan Kaufmann Publishers Inc.

References IV

[Shwartz et al., 2017] Shwartz, V., Stanovsky, G., and Dagan, I. (2017). Acquiring predicate paraphrases from news tweets. In **SEM*, pages 155–160.

[Stanovsky and Dagan, 2016] Stanovsky, G. and Dagan, I. (2016). Annotating and predicting non-restrictive noun phrase modifications. In *ACL*.

[Stanovsky et al., 2016] Stanovsky, G., Ficler, J., Dagan, I., and Goldberg, Y. (2016). Getting more out of syntax with props. *arXiv*.

[Williams et al., 2018] Williams, A., Nangia, N., and Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, New Orleans, Louisiana.