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

July 18, 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 $\Rightarrow$ 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 $\Rightarrow$ 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 $\Rightarrow$ ENTAILMENT
2. A juggler entertaining a group of children on the street $\Rightarrow$ 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

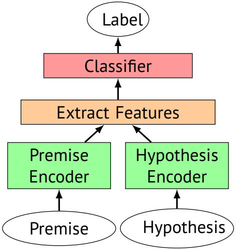
Hypotheses

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

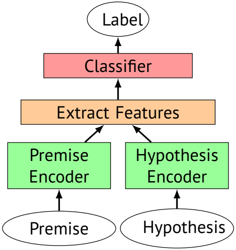
- End-to-end, either **sentence-encoding** or **attention-based**

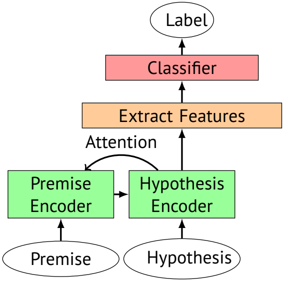


1

Neural NLI Models

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

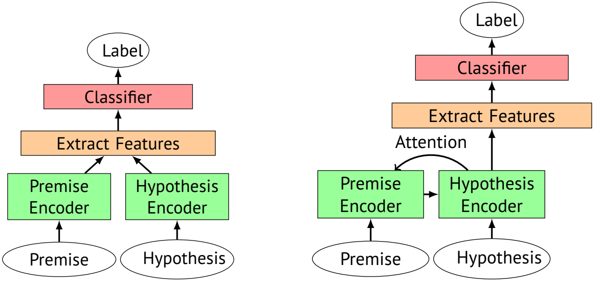




1

Neural NLI Models

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

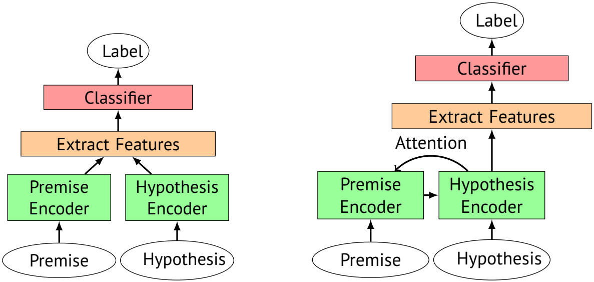


- Lexical knowledge: only from pre-trained word embeddings
  - As opposed to using resources like WordNet

1

Neural NLI Models

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

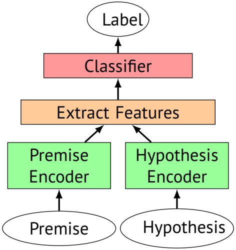


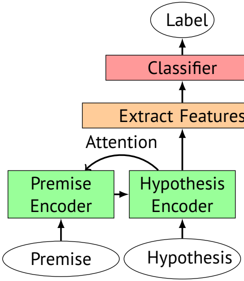
- Lexical knowledge: only from pre-trained word embeddings
  - As opposed to using resources like WordNet
- SOTA exceeds human performance...

1

Neural NLI Models

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





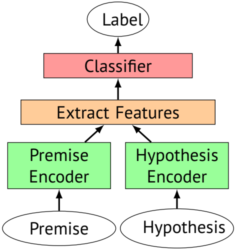
- Lexical knowledge: only from pre-trained word embeddings
  - As opposed to using resources like WordNet
- SOTA exceeds human performance... [1]

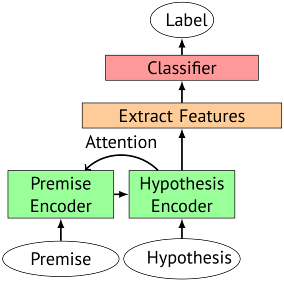
---

[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"

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'

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'
  - w' is in the SNLI vocabulary and in pre-trained embeddings
  - Crowdsourcing labels (mostly contradictions!)

Contradiction

The man is holding a <u>saxophone</u> $\rightarrow$ The man is holding an <u>electric guitar</u>

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 <u>saxophone</u> $\rightarrow$ The man is holding an <u>electric guitar</u>

Entailment

A little girl is very <u>sad</u> $\rightarrow$ A little girl is very <u>unhappy</u>

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 <u>saxophone</u> $\rightarrow$ The man is holding an <u>electric guitar</u>

Entailment

A little girl is very <u>sad</u> $\rightarrow$ A little girl is very <u>unhappy</u>

Neutral

A couple drinking <u>wine</u> $\rightarrow$ A couple drinking <u>champagne</u>

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]

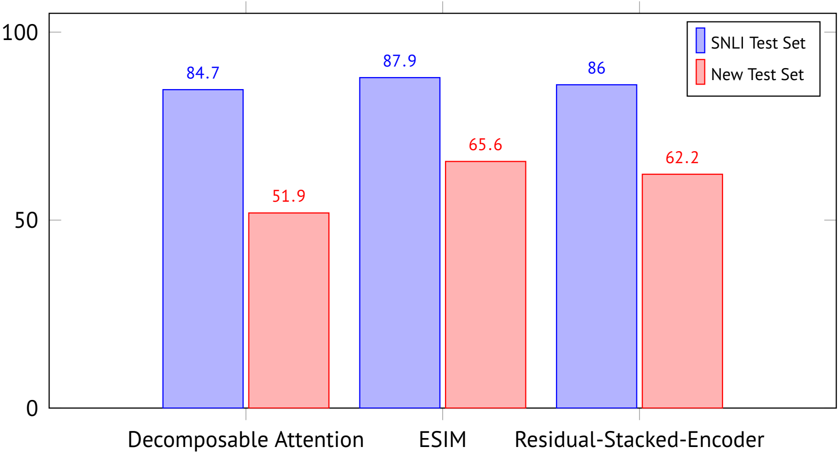
Evaluation Setting

- 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

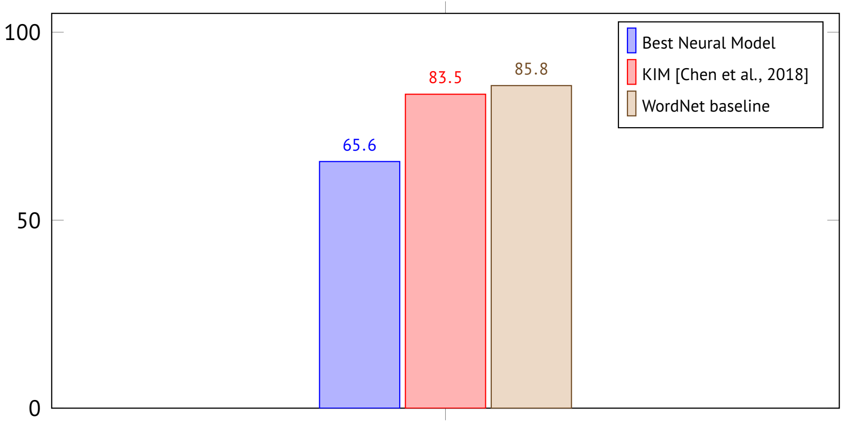
Can neural NLI models recognize lexical inferences?



Dramatic drop in performance across models.

Sanity Check

Performance of WordNet-informed Models



The test set is solvable using WordNet.

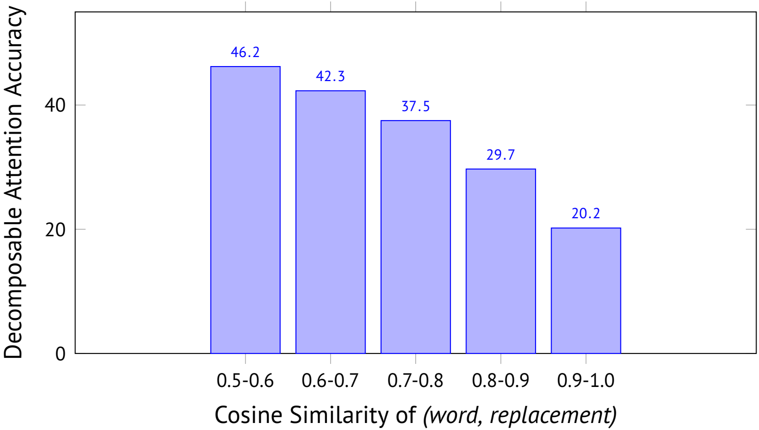
What do neural NLI models learn with respect to lexical semantic relations?

Analysis 1: Word Similarity

- Models err on contradicting word-pairs with similar embeddings
  - *A man starts his day in India* $\rightarrow$ *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

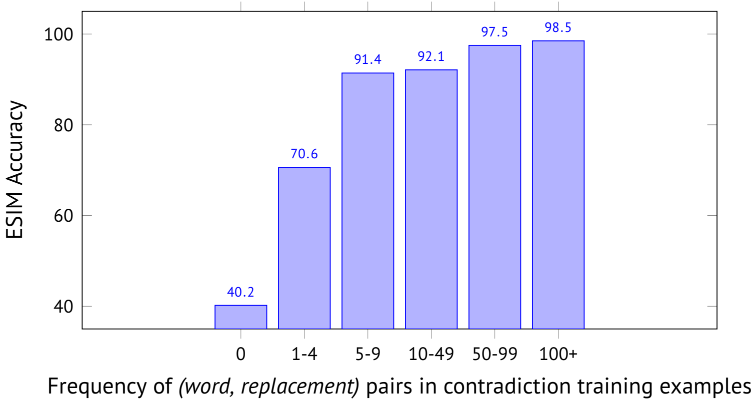


Analysis 2: Frequency in Training

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

Analysis 2: Frequency in Training

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



Recap

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

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

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

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
- May be used as a complementary test set to assess the lexical inference abilities of NLI systems

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

- May be used as a complementary test set to assess the lexical inference abilities of NLI systems

Thank you!

References

[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.

[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.

[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.

[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)*.