A large annotated corpus for learning natural language inference

Samuel R. Bowman, Gabor Angeli,
Christopher Potts, Christopher D. Manning

Stanford NLP Group
Stanford University

EMNLP 2015

Outline

Natural Language Inference (NLI)

- a.k.a. recognising textual entailment (RTE)
- Does a piece of text follows from or contradict another?

Example

Premise: A man inspects the uniform of a figure in some East Asian country.
Hypothesis: The man is sleeping.

Example

Premise: A soccer game with multiple males playing.
Hypothesis: Some men are playing a sport.

Natural Language Inference (NLI)

- a.k.a. recognising textual entailment (RTE)
- Does a piece of text follows from or contradict another?

Example

Premise: A man inspects the uniform of a figure in some East Asian country.
Hypothesis: The man is sleeping.
contradiction

Example

Premise: A soccer game with multiple males playing.
Hypothesis: Some men are playing a sport.
entailment

SNLI Corpus

- Bowman et al. 2015
- 570,152 pairs of sentences
- Collected over Amazon Mechanical Turk
  - Simple annotation guidelines
  - Captions from the Flickr30k corpus
  - 2,500 workers contributed
  - 30 trusted workers for validation
- https://nlp.stanford.edu/projects/snli/

Outline

1. Natural Language Inference
2. Stanford Natural Language Inference Corpus
   - Coreference issue
   - Data collection & validation
3. Methods
   - Feature-based
   - Sentence embedding
4. Results

Example

A boat sank in the Pacific Ocean.
A boat sank in the Atlantic Ocean.

Do they contradict each other? or neutral?

SNLI Corpus

Coreference issue

Example

A boat sank in the Pacific Ocean.
A boat sank in the Atlantic Ocean.

Do they contradict each other? or neutral?

Outline

SNLI Corpus

Data collection

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption **An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.**

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption. **Entailment**

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be. **Neutral**

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption. **Contradiction**

Problems (optional) *If something is wrong, have a look at the FAQ, do your best above, and let us know here.*

Figure: Crowd-sourcing on Amazon Mechanical Turk[1]

[1] https://www.nyu.edu/projects/bowman/cusp_snli_slides.pdf

SNLI Corpus

Data collection

Data set sizes:	
Training pairs	550,152
Development pairs	10,000
Test pairs	10,000
Sentence length:	
Premise mean token count	14.1
Hypothesis mean token count	8.3
Parser output:	
Premise 'S'-rooted parses	74.0%
Hypothesis 'S'-rooted parses	88.9%
Distinct words (ignoring case)	37,026

Figure: Key statistics for the raw sentence pairs in SNLI

Condition	% of pairs
5 vote unanimous agreement:	58.3%
3-4 vote majority for one label including author:	32.9%
3-4 vote majority for one label not including original author:	6.8%
No majority for any one label:	2.0%

Figure: Data validation with 10% held-out data[2]

[2] https://www.nyu.edu/projects/bowman/cusp_snli_slides.pdf

Outline

1. Natural Language Inference
2. Stanford Natural Language Inference Corpus
   - Coreference issue
   - Data collection & validation
3. Methods
   - Feature-based
   - Sentence embedding
4. Results

Feature-based

Feature selection

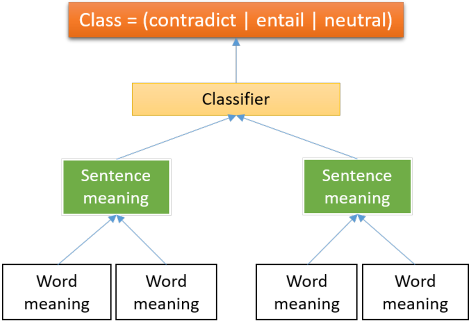
1. **BLEU score** of the hypothesis with respect to the premise (n-gram =1..4)
2. **Length difference** between the hypothesis and the premise
3. **Count & percentage of overlapping words** in the premise and hypothesis (all and just nouns, verbs, adjectives, and adverbs)
4. **Unigram and bigram** in the hypothesis.
5. **Cross-unigrams**: for every pair of words across the premise and hypothesis which share a POS tag.
6. **Cross-bigrams**: for every pair of bigrams across the premise and hypothesis which share a POS tag on the second word.

Outline

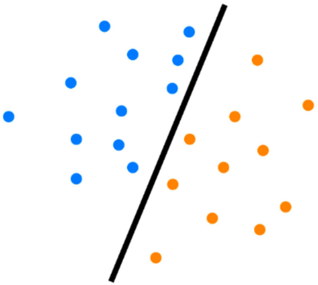
- Represent words as a vectors
  - *cat* $\Rightarrow [0.4, 0.3, ..., 0.1]$
  - *dog* $\Rightarrow [0.1, 0.5, ..., 0.2]$
- Word2vec Mikolov et al. 2013

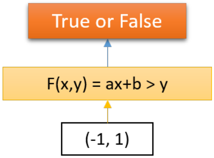
Sentence embedding

Bowman et al. 2015



Learning a function





Sentence embedding

Bowman et al. 2015

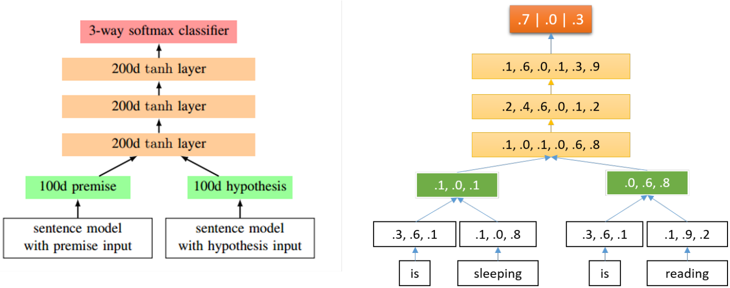


Figure: Neural network classification architecture

Sentence embedding

Bowman et al. 2015

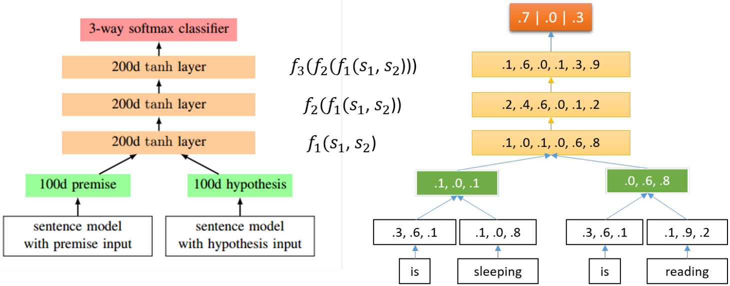


Figure: Neural network classification architecture

Results

Feature-based

System	SNLI		SICK	
	Train	Test	Train	Test
Lexicalized	99.7	**78.2**	90.4	**77.8**
Unigrams Only	93.1	71.6	88.1	77.0
Unlexicalized	49.4	50.4	69.9	69.6

Figure: Feature-based variants

Results

Feature-based vs. Sentence embedding

Model	% Accuracy (Test set)
Feature-based classifier	78.2
LSTM RNN sequence model	80.6

Figure: Results[3]

[3] https://nlp.stanford.edu/manning/talks/SIGIR2016-Deep-Learning-NLI.pdf

Summary

- A large-scale, naturalistic corpus of sentence pairs labeled for entailment, contradiction, and independence.
- Both simple lexicalised models and neural network models perform well.
- The RepEval 2017 Shared Task
  - https://repeval2017.github.io/shared/

References I

Bowman, Samuel R. et al. (2015). "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Mikolov, Tomas et al. (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.