A large annotated corpus of entailments and contradictions

Samuel R. Bowman
Gabor Angeli
Christopher Potts
Christopher D. Manning
Stanford University

Premise: *A man speaking or singing into a microphone while playing the piano.*
Hypothesis: *A man is performing surgery on a giraffe while singing.*
Label: contradiction

Corpus Semantics and Entailments

Corpora for semantics?

Answering questions about what sentences people are willing or able to say:

- Almost any corpus of naturally occurring text, like...
  - CoCA
  - BNC
  - The Web
  - …

Corpora for semantics?

Answering questions about how people interpret particular structures:

- Collecting new human judgments
- Corpora annotated with meanings:
  (sentence, context, meaning)

Semantic annotations

- How should the meaning be represented?
  - Metalanguage sentences
    - Logical forms:
      - Catch 22.
  - Nonlinguistic data:
    - Images or videos
      - Difficult to study statistically
    - States of affairs in a game (cf. Potts, 2012)
      - Very limited domain
  - Object language sentences:
    - Natural language inference corpora!

Natural language inference (NLI)

also known as recognizing textual entailment (RTE)

Premise: *James Byron Dean refused to move without blue jeans*

Hypothesis: *James Dean didn't dance without pants*

Label: {**entails**, contradicts, neither}

Example: MacCartney thesis '09

NLI in Semantics

Logical forms make predictions about entailments.

If *blue_jeans* = λx . pants(x) $\wedge$ denim(x) $\wedge$ blue(x)

and *pants* = λx . pants(x)

then *The dog is wearing blue jeans.*

entails *The dog is wearing pants.*

...assuming reasonable logical forms for the rest of the words in the sentence.

cf. research Natural Logic: A logic that isolates entailment predictions without using explicit logical forms for sentences.

NLI in Semantics

Correctly predicting entailments requires capturing every aspect of semantics except grounding:

- Lexical entailment:
  - Does *jeans* entail *pants*?
- Quantification:
  - Does *most* N V entail *some* N V?
- Factivity and implicativity
  - Does N *managed to* V entail N V*ed*?
- Object and event coreference
- Lexical ambiguity and scope ambiguity
- Propositional attitudes
- Modality

...

Existing NLI Corpora

Natural language inference data

Corpus	Size	Full Sentences	Expert Curated	Model Biased
FraCaS	.3k	✓	×	
RTE	7k	✓	×	
SICK	10k	✓	×	
DG	728k	~		×
Levy	1,500k			×
PPDB	100,000k			×

Natural language inference data

Corpus	Size	Full Sentences	Expert Curated	Model Biased
FraCaS	.3k	✓	×	
RTE	7k	✓	×	
SICK	10k	✓	×	
SNLI	570k	✓		
DG	728k	~		×
Levy	1,500k			×
PPDB	100,000k			×

The Stanford NLI Corpus

- The Stanford Natural Language Inference Corpus (SNLI)
- 570k pairs of sentences
- Collected over Amazon Mechanical Turk
  - Simple annotation guidelines
  - No model necessary

Coreference issues for NLI data collection

One event or two?

Premise: A boat sank in the Pacific Ocean.

Hypothesis: A boat sank in the Atlantic Ocean.

One event or two? One.

Premise: A boat sank in the Pacific Ocean.

Hypothesis: A boat sank in the Atlantic Ocean.

Label: Contradiction

One event or two? One.

Premise: Ruth Bader Ginsburg was appointed to the US Supreme Court.

Hypothesis: I had a sandwich for lunch today

Label: Contradiction

One event or two? Two.

Premise: Ruth Bader Ginsburg was appointed to the US Supreme Court.

Hypothesis: I had a sandwich for lunch today

Label: Neutral

One event or two? Two.

Premise: A boat sank in the Pacific Ocean.

Hypothesis: A boat sank in the Atlantic Ocean.

Label: Neutral (two different events)

No contradictions at all in typical natural sentences.

Our data collection prompt

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.

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.

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.

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

Source captions from Flickr30k: Young, Lai, Hodosh, and Hockenmaier, TACL '14

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.

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.

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

Source captions from Flickr30k: Young, Lai, Hodosh, and Hockenmaier, TACL '14

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.

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

Source captions from Flickr30k: Young, Lai, Hodosh, and Hockenmaier, TACL '14

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.*

Source captions from Flickr30k: Young, Lai, Hodosh, and Hockenmaier, TACL '14

One *photo* or two? One.

Premise: Ruth Bader Ginsburg being appointed to the US Supreme Court.

Hypothesis: A man eating a sandwich for lunch.

Label: Different photo (⊢ contradiction)

Flickr30k descriptive captions

The source of the premise in almost every* SNLI pair:

Premise: A white dog with long hair jumps to catch a red and green toy.

Hypothesis: An animal is jumping to catch an object.

Label: Entailment

Flickr30k: Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier, TACL ‘14

- Collected over Mechanical Turk

*A small subset of the data draws on visualgenome.org data from a pilot study.

What we got

Some sample results

Premise: Two women are embracing while holding to go packages.

Entailment: Two woman are holding packages.

Some sample results

Premise: A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.

Neutral: A man is repainting a garage

Some sample results

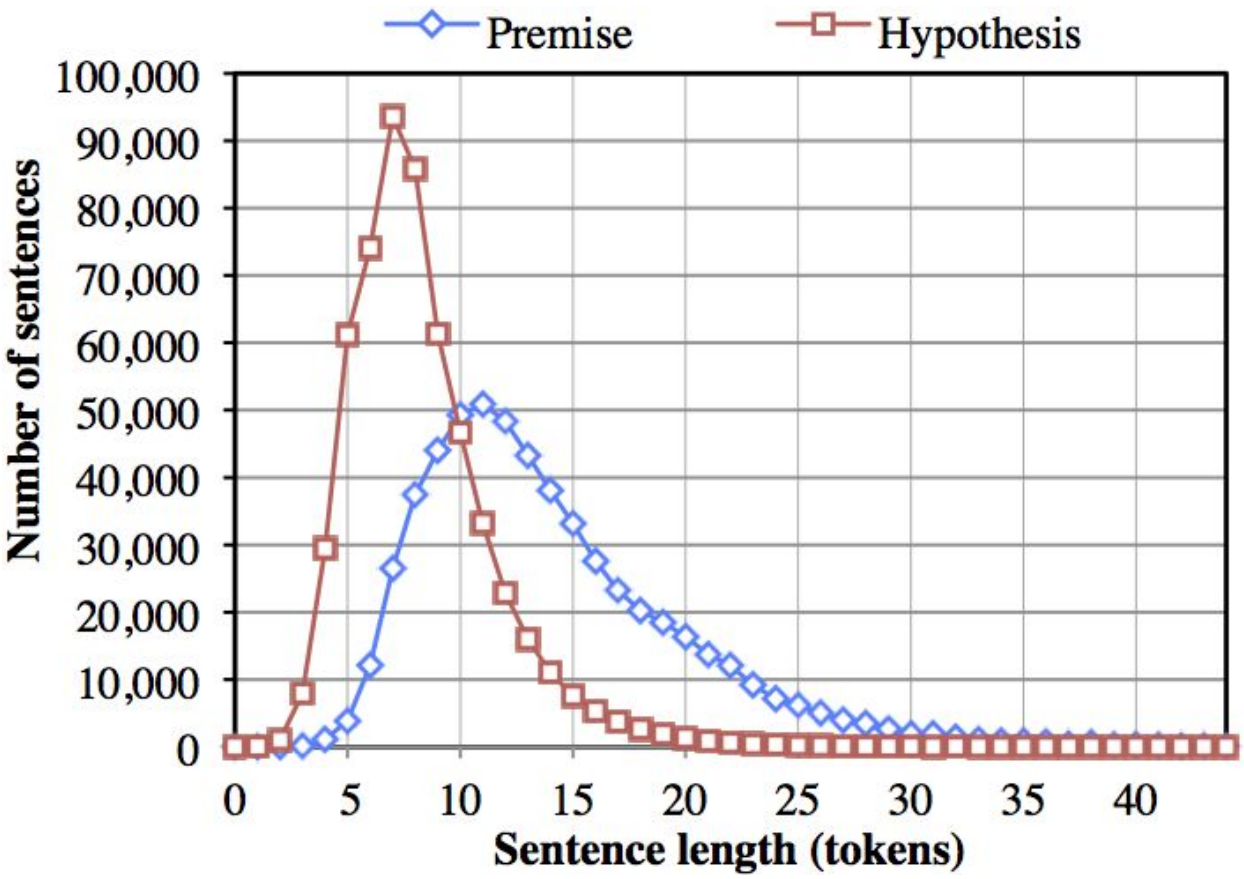
Premise: A man selling donuts to a customer during a world exhibition event held in the city of Angeles

Contradiction: A woman drinks her coffee in a small cafe.

General observations

- Entailments aren't logically 'strict', and incorporate commonsense background assumptions.
- Only see semantic phenomena which can occur in scene descriptions. (Hard to avoid.)
- Premise and hypothesis sentences can be syntactically (and stylistically) very different.
- Spelling and grammar errors are rare.
  (NB: the corpus is not cleaned)
- Contradiction hypotheses are often (but not always) somehow related to their premises.

Sentence length



Bare NPs vs. full sentences

Data source	**% full sentences** (Stanford Parser 'S')
Premises (Flickr30k)	74.0
Hypotheses (our work)	88.9

Data validation

Data validation

Premise: Two women are embracing while holding to go packages.

Entailment: Two woman are holding packages.

Data validation

Premise: Two women are embracing while holding to go packages.

Entailment: Two woman are holding packages.

Labels: entailment

Data validation

Premise: Two women are embracing while holding to go packages.

Entailment: Two woman are holding packages.

Labels: entailment entailment entailment neutral entailment

Data validation

Premise: Two women are embracing while holding to go packages.

Entailment: Two woman are holding packages.

Labels: entailment entailment entailment neutral entailment

Gold label: entailment

Data validation

Premise: Two women are embracing while holding to go packages.

Entailment: Two woman are holding packages.

Labels: entailment entailment entailment neutral entailment

Gold label: entailment

Relabeled 10% of SNLI.

Instructions

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

Your job is to figure out, based on the correct caption for a photo, if another caption is also correct:

- Choose **definitely correct** if any photo that was captioned with the caption on the left would also fit the caption on the right. Example: *"A kitten with spots is playing with yarn."/"A cat is playing."*
- Choose **maybe correct** if the second caption could describe photos that fit the first caption, but could also describe sentences that don't fit the first caption. Example: *"A kitten with spots is playing with yarn."/"A kitten is playing with yarn on a sofa."*
- Choose **definitely incorrect** if any photo that could possibly be captioned with the caption on the left would not fit the caption on the right. Example: *"A kitten with spots is playing with yarn."/"A puppy is playing with yarn."*

We have already labeled one out of every 250 HITs. Completeing one of these HITs yields a bonus of $1 for each response that matches our label for up to $5. More questions? See the FAQ.

Correct caption	Candidate caption	Def. correct	Maybe correct	Def. incorrect
A girl in a hat steers her electric wheelchair.	A girl is running outside.	○	○	○
This is a picture of three young men, dressed in suits with one on a bike, and of a young lady wearing a white dress.	a dog sleeps	○	○	○
A man in a purple and blue shirt and shorts is getting ready to hit a golf ball.	A man is on a golf course.	○	○	○
A man is making a phone call.	A man is calling his wife.	○	○	○
Two girls are playing in the snow and throwing snowballs.	The two girls have no arms.	○	○	○

Problems (optional) *If something is wrong with a caption that makes it hard to understand (more than just a typo), do your best above and let us know here.*

The results

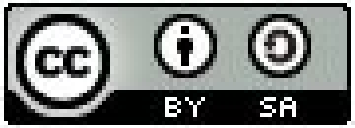
Condition	% of pairs
5 vote unanimous agreement:	58.3%
3-4 vote majority for one label including author:	32.9%
3-4 vote majority for one label not including original author:	6.8%
No majority for any one label:	2.0%

Give it a try!

Available now (with an accompanying paper):

nlp.stanford.edu/projects/snli

Download: SNLI 1.0 (zip, ~100MB)



- Standard distribution includes JSON and tab-separated text.
- Sentences are included both raw and tokenized+parsed with Stanford CoreNLP.

Thanks!

Download the corpus:

nlp.stanford.edu/projects/snli

More questions?

sbowman@stanford.edu

We gratefully acknowledge support from a Google Faculty Research Award, a gift from Bloomberg L.P., the Defense Advanced Research Projects Agency (DARPA) Deep Exploration and Filtering of Text (DEFT) Program under Air Force Research Laboratory (AFRL) contract no. FA8750-13-2-0040, the National Science Foundation under grant no. IIS 1159679, and the Department of the Navy, Office of Naval Research, under grant no. N00014-10-1-0109. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Google, Bloomberg L.P., DARPA, AFRL NSF, ONR, or the US government. *Our MTurk workers were amazing, and we thank them too.*

Validation: Other metrics

% of total labels matching gold label	89.0%
% of total labels matching author's label	85.8%

Fleiss κ	
entailment	0.72
contradiction	0.77
neutral	0.60
overall	0.70

Realities of Mechanical Turk

- Employed ~2,500 workers
  - Used several worker qualification strategies, and bonuses in validation
- Reviewed and discarded ~100 pairs that were marked as problematic (either in collection or validation).
  - Data entry errors (early sumbission)
  - Bad source captions from Flickr30k
    ("The image didn't load")
  - Uninterpretable English
- Updated FAQ ~10 times to discourage overly regular data
- Caught ~20 cases of fraud
  - Mostly random guessing on the validation task
- Reddit (HWTF) reviews extremely positive: Many found task fun