A large annotated corpus for learning natural language inference

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

Presenter: Medhini G Narasimhan

Outline

- Entailment and Contradiction
- Examples of Natural Language Inference
- Prior datasets for Natural Language Inference
- Shortcomings of previous work
- Stanford Natural Language Inference Corpus
- Data Collection
- Data Validation
- Models on this dataset
- Conclusion

Entailment and Contradiction

- **Entailment**: The truth of one sentence implies the truth of the other sentence.

  "It is raining heavily outside."

  entails

  "The streets are flooded."

- **Contradiction**: The truth of one sentence implies the falseness of the other.

  "It is cold in here."

  contradicts

  "It is hot in here."

- Understanding entailment and contradiction is fundamental to understanding natural language.
- **Natural Language Inference:** Determining whether a natural language hypothesis can justifiably be inferred from a natural language premise.

Examples of Natural Language Inference

Neutral

A woman with a green headscarf, blue shirt and a very big grin.

The woman is young.

Entailment

A land rover is being driven across a river.

A Land Rover is splashing water as it crosses a river.

Contradiction

An old man with a package poses in front of an advertisement.

A man walks by an ad.

Objective

To introduce a Natural Language Inference corpus which would allow for the development of improved models on entailment and contradiction and Natural Language Inference as a whole.

Prior datasets for NLI

- **Recognizing Textual Entailment(RTE)** challenge tasks:
  - High-quality, hand-labelled data sets.
  - Small in size and complex examples.
- **Sentences Involving Compositional Knowledge (SICK)** data for the SemEval 2014:
  - 4,500 training examples.
  - Partly automatic construction introduced some spurious patterns into the data.
- **Denotation Graph** entailment set:
  - Contains millions of examples of entailments between sentences and artificially constructed short phrases.
  - Labelled using fully automatic methods, hence noisy.

Issues with previous datasets

- Too small in size to train modern data-intensive wide-coverage models.
- Indeterminacies of event and entity coreference lead to indeterminacy concerning the semantic label.
- Event indeterminacy:
    - *A boat sank in the Pacific Ocean* and *A boat sank in the Atlantic Ocean.*
    - Contradiction if they refer to the same event, else neutral.
- Entity indeterminacy:
    - *A tourist visited New York* and *A tourist visited the city.*
    - If we assume coreference, this is entailment, else neutral.

Stanford Natural Language Inference corpus

- Freely available collection of 570K labelled sentence pairs, written by humans doing a novel grounded task based on image captioning.
- The labels include **entailment**, **contradiction**, and **semantic independence**.
- Image captions would ground examples to specific scenarios and overcome entity and event indeterminacy.
- Participants allowed to produce entirely novel sentences which led to richer examples.
- A subset of the resulting sentences were sent to a validation task in order to provide a highly reliable set of annotations.

Data Collection

- Premises obtained from Flickr30K image captioning dataset.
- Using just the captions, workers were asked to generate entailing, neutral and contradictive examples.

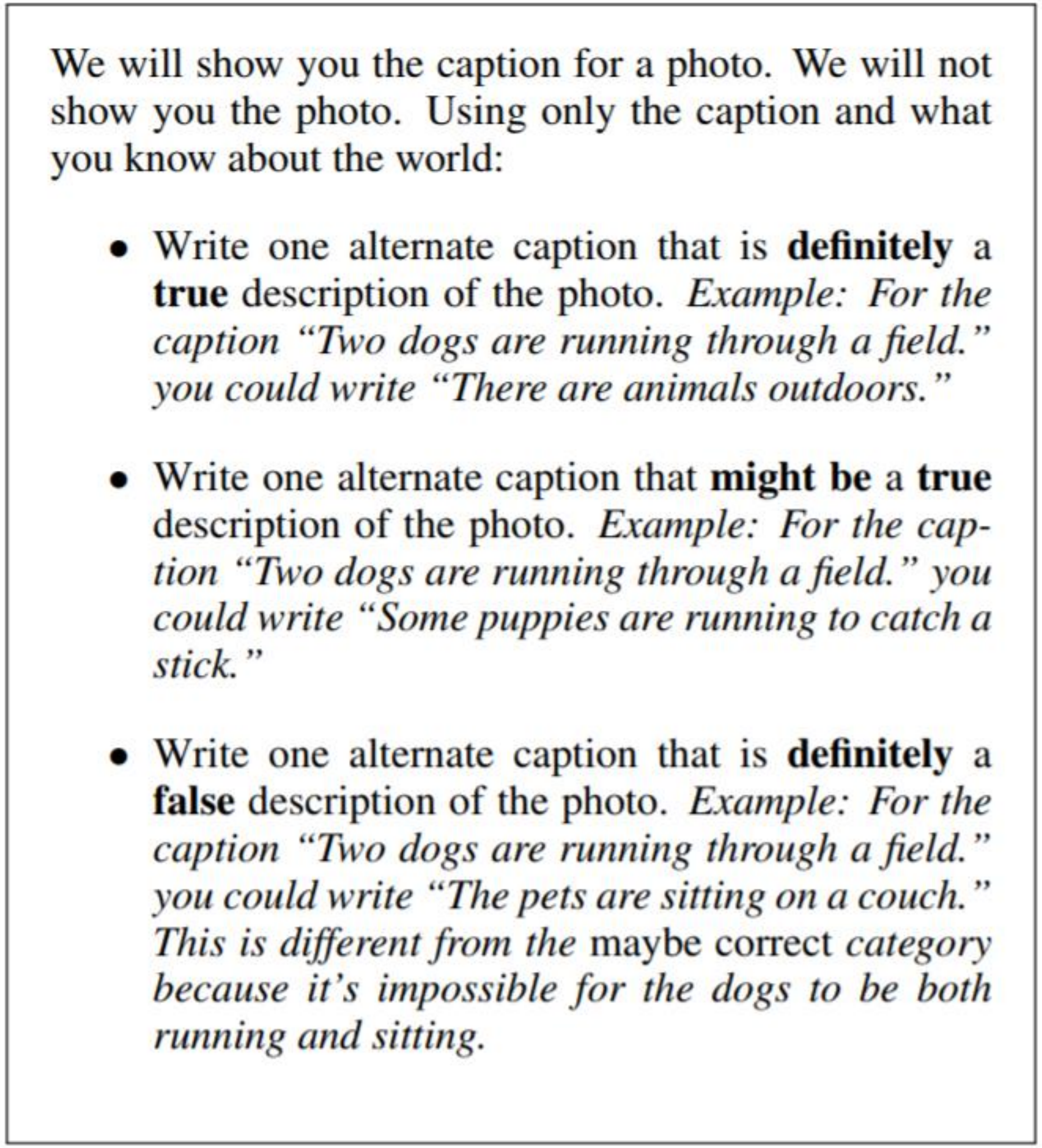
A female tennis player in a purple top and black skirt swings her racquet.
A female tennis player preparing to serve the ball.
A woman in a purple tank top holds a tennis racket, extends an arm upward, and looks up.
A woman wearing a purple shirt and holding a tennis racket in her hand is looking up.
Girl is waiting for the ball to come down as she plays tennis.

A man is snow boarding and jumping off of a snow hill.
A person in a black jacket is snowboarding during the evening.
A silhouette of a person snowboarding through a pile of snow.
A snowboarder flying off a snow drift with a colourful sky in the background.
The person in the parka is on a snow board.

A motorcycle races.
A motorcycle rider in a white helmet leans into a curve on a rural road.
A motorcycle rider making a turn.
Someone on a motorcycle leaning into a turn.
There is a professional motorcyclist turning a corner.

Data Collection

- The sentences in SNLI are all descriptions of scenes, and photo captions.
- Reliable judgments from untrained annotators
- Logically consistent definition of *contradiction*.
- Issues of coreference greatly mitigated. For example, "A dog is lying in the grass", the main object is the dog.



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:

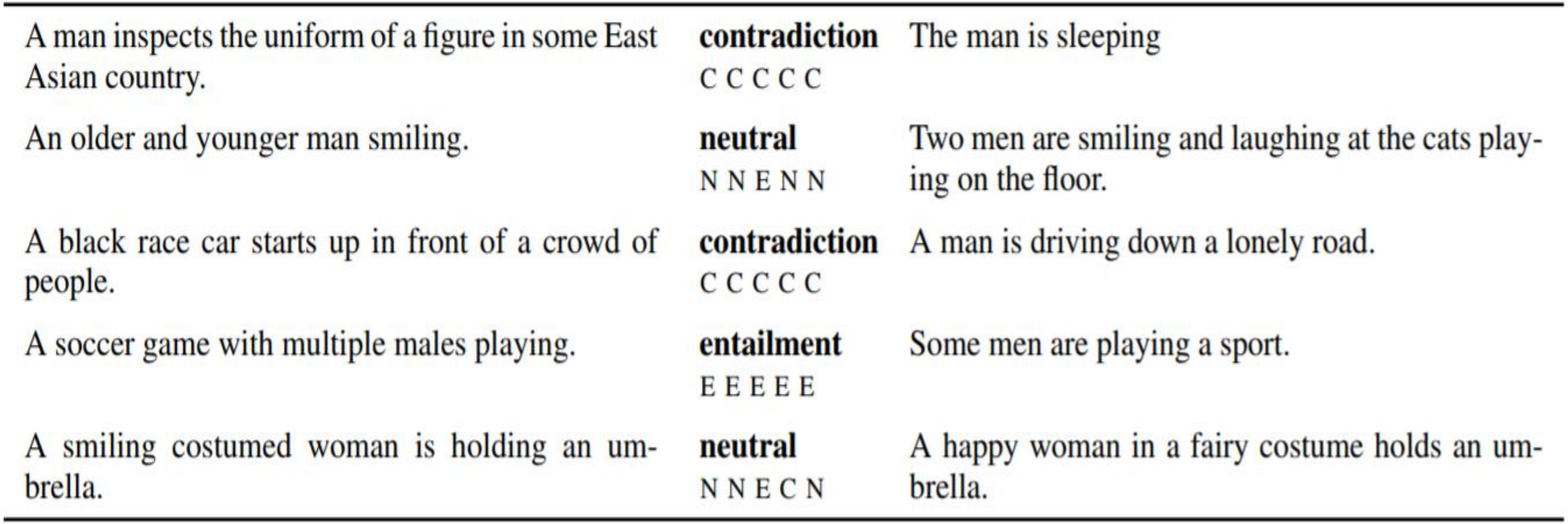
- Write one alternate caption that is **definitely** a **true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "There are animals outdoors."*
- Write one alternate caption that **might be** a **true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "Some puppies are running to catch a stick."*
- Write one alternate caption that is **definitely** a **false** description of the photo. *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.*

Figure 1: The instructions used on Mechanical Turk for data collection.

Data Validation

- Measure the quality of corpus and collect additional data for test and development sets.
- Validation is done by asking four annotators to label the same pair, this gave five labels per pair.
- Based on their labelling skills, 30 trusted workers were picked.
- Sentence pair assigned a gold label if one of the three labels were chosen by at least three of the five annotators.
- Only sentence pairs with gold label used during model building.

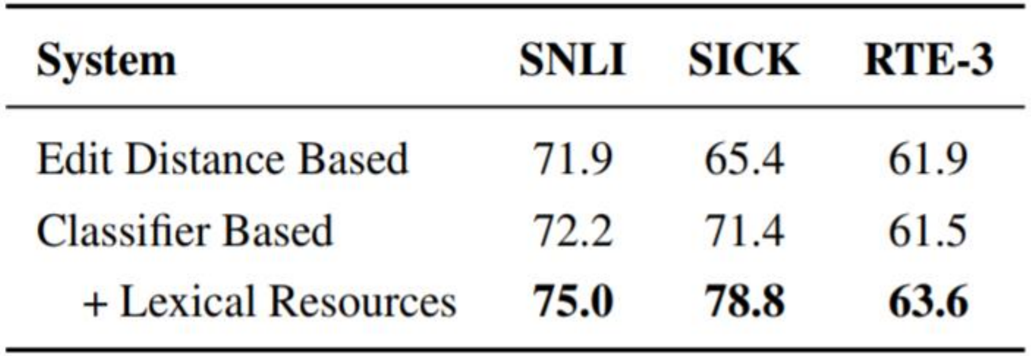
Stanford Natural Language Inference corpus



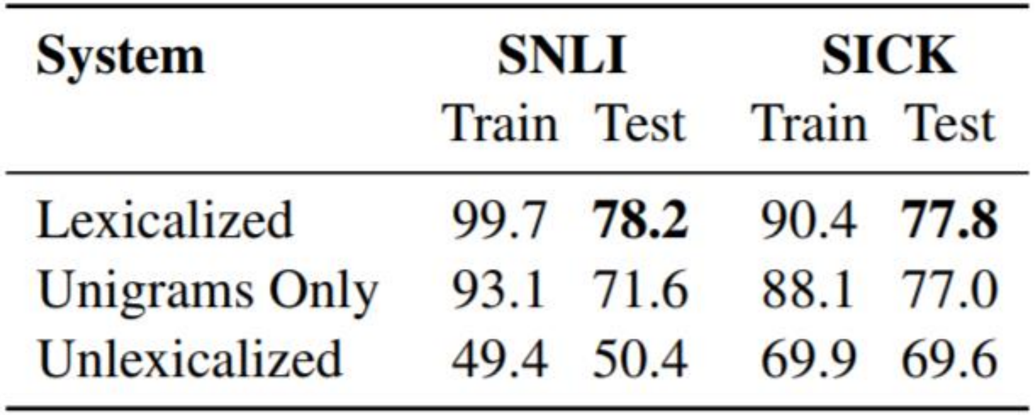
A man inspects the uniform of a figure in some East Asian country.	**contradiction** C C C C C	The man is sleeping
An older and younger man smiling.	**neutral** N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	**contradiction** C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	**entailment** E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	**neutral** N N E C N	A happy woman in a fairy costume holds an umbrella.

Models and Results on SNLI

- Excitement Open Platform Model
  - Edit distance algorithm: Tunes the weight of the three case insensitive edit distance operations.
  - Simple lexical based classifier.
- Lexicalized feature-based classifier model
  - BLEU Score.
  - Length difference.
  - Overlap between words.
  - Indicator for every unigram and bigram.
  - Cross unigrams.
  - Cross bigrams.



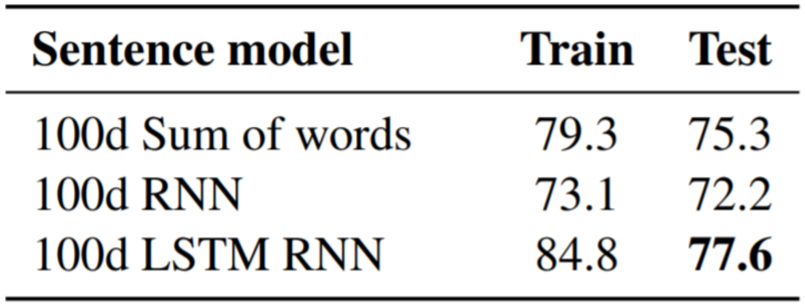
System	SNLI	SICK	RTE-3
Edit Distance Based	71.9	65.4	61.9
Classifier Based	72.2	71.4	61.5
+ Lexical Resources	**75.0**	**78.8**	**63.6**



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

Models and Results on SNLI

- Neural network sequence model
  - Generate vector embedding of each sentence.
  - Train classifier to label the vectors.
  - Two sequence embedding models: Plan RNN and LSTM RNN.
  - Embeddings initialized with GloVE vectors.
  - Lexicalized model performs better.



Sentence model	Train	Test
100d Sum of words	79.3	75.3
100d RNN	73.1	72.2
100d LSTM RNN	84.8	**77.6**

Table 6: Accuracy in 3-class classification on our training and test sets for each model.

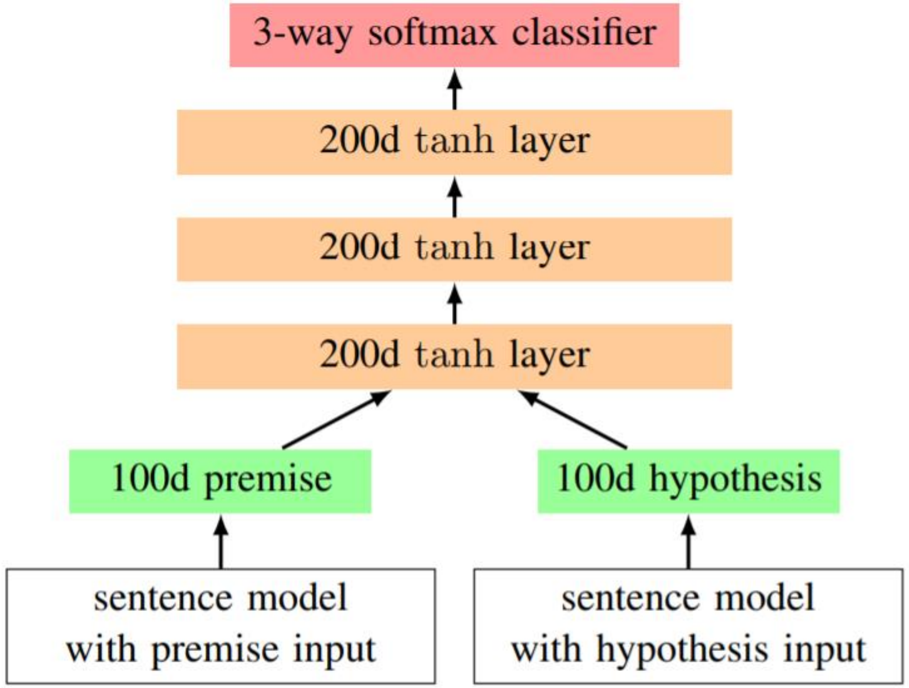


Figure 3: The neural network classification architecture: for each sentence embedding model evaluated in Tables 6 and 7, two identical copies of the model are run with the two sentences as input, and their outputs are used as the two 100d inputs shown here.

Conclusion

- SNLI draws fairly extensively on common sense knowledge.
- Hypothesis and premise sentences often differ structurally in significant ways.
- Sentences collected are largely fluent, correctly spelled English.
- Basic models were introduced which have been outperformed.
- Future directions – Using entailment and contradiction pairs to generate question answers on Flickr30k.

Questions?

Thank You!