A large annotated corpus for learning natural language inference

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

Presented by: Filip Trhlik

What is Natural language inference (NLI)?

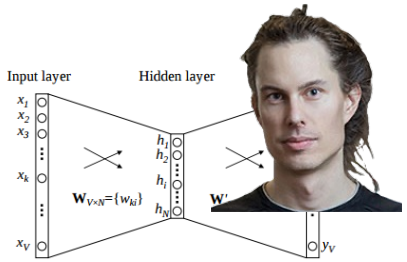
Recognising language entailment is a fundamental NLP task[1]:

- Model is given premise and hypothesis to identify their relation
- **Entailment** / *Neutral* / Contradiction
- Information retrieval, semantic parsing, commonsense reasoning...

👉 Filip is giving a presentation (E)
👉 Sandwich is good (N)
👉 Filip is bungee jumping (C)

[1] Jerrold J. Katz. 1972. Semantic Theory. Harper & Row, New York.

Positioning the paper in the history



2012[2] 2013[3] 2015 2017[4] 2022

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun.
[3] Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations.
[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems

Goal of this paper

- In the past, NLI models leveraged a variety of techniques leveraging predefined features
- This paper asks whether **domain-general neural approaches to NLI using distributed representations are sufficient** and whether such models can learn sentence meaning (that is, make the necessary logical and commonsense inferences).
- In order to do this, it seeks to confront the **data problem**

Data problem

- A large amount of good labeled data is needed to train these models (they did not have pre-training)
- Previous NLI corpora does not support such training:

Corpus	Size	Natural	Validated
FraCaS	.3k	∼	✓
RTE	7k	✓	✓
SICK	10k	✓	✓
DG	728k	∼	
Levy	1,500k		
PPDB	100,000k	∼	

Stanford NLI dataset

- Authors provide a new NLI dataset with **570,152 sentence pairs**
- All sentences & labels written by humans in a grounded, naturalistic setup
- 50-times larger than the previous SOTA dataset

Annotation challenges

- Indeterminacy of event and entity (one or two events)[5]

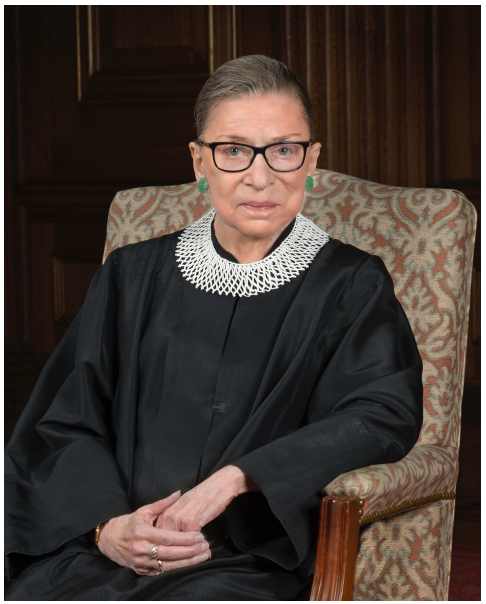
👉A boat sank in the Pacific Ocean
👉A boat sank in the Atlantic Ocean.

[5] Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. 2008. Finding contradictions in text. In Proc. ACL.

Annotation challenges

- Indeterminacy of event and entity (one or two events)

👉Ruth Bader Ginsburg was appointed to the US Supreme Court
👉I had a sandwich for lunch today

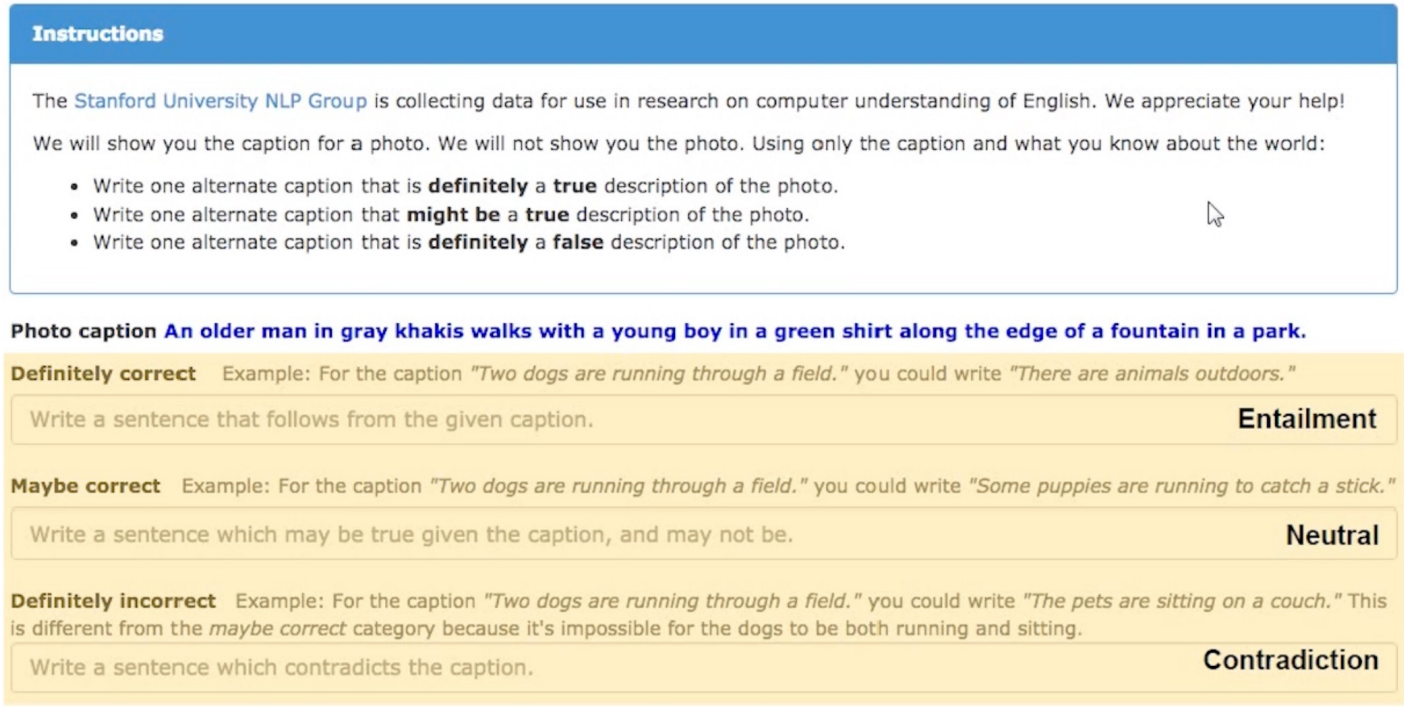


Annotation challenges

- Indeterminacy of event and entity (one or two events)

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

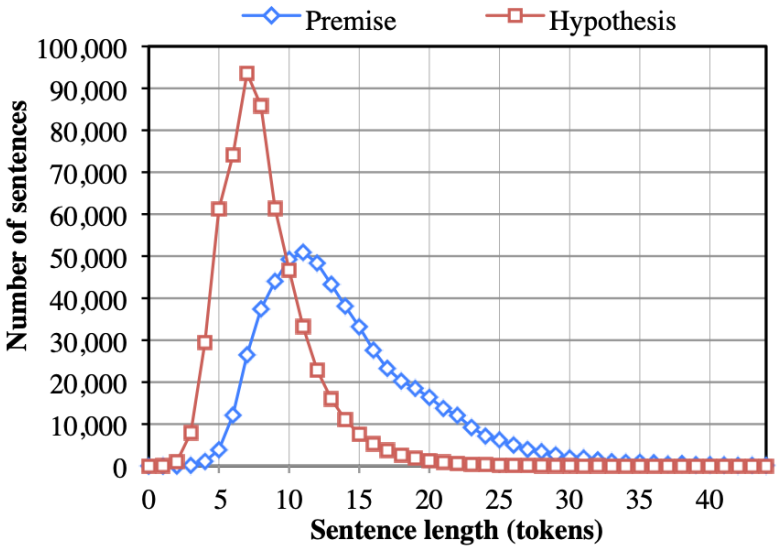
Annotation challenges



👉 Collected through Amazon Mechanical Turk

Data

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



Data validation

- Four extra judgments for 10% of pairs (5 labels each incl. author)

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

- Fleiss κ[6]: contradiction 0.77, entailment 0.72, neutral 0.60, overall 0.70.

[6] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382.

Standard classifiers

- Variants of a simple feature-based classifier model, which makes use of both unlexicalized and lexicalized features

Features:

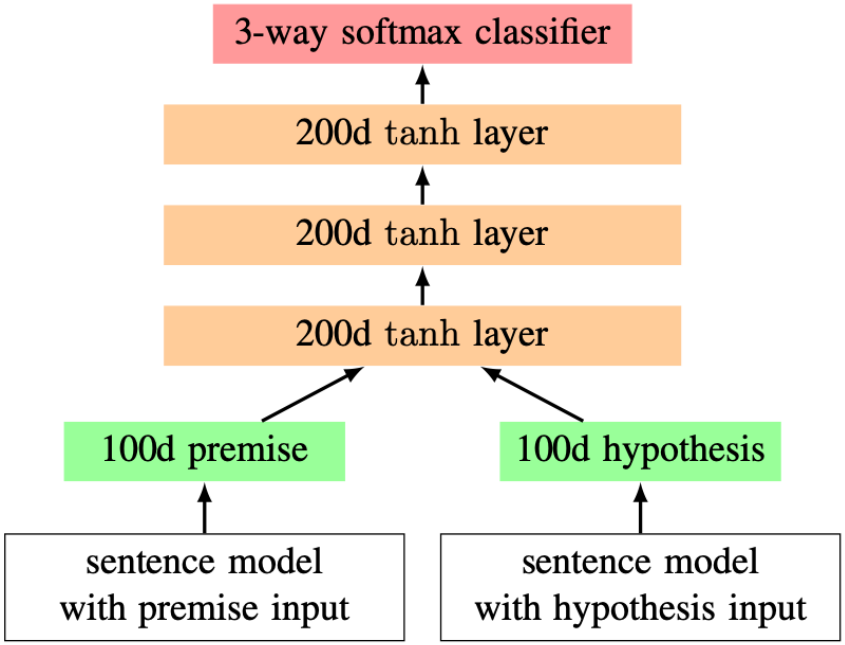
- BLEU; length diff; word/PoS overlap; lexical features (unigrams & bigrams); cross-unigrams and cross-bigrams.

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

NLI Neural Experiment

- Authors utilised sentence embedding as an intermediate steps for NLI
- **Classifier:** 3×200-d tanh stack over concatenated embeddings + softmax

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**

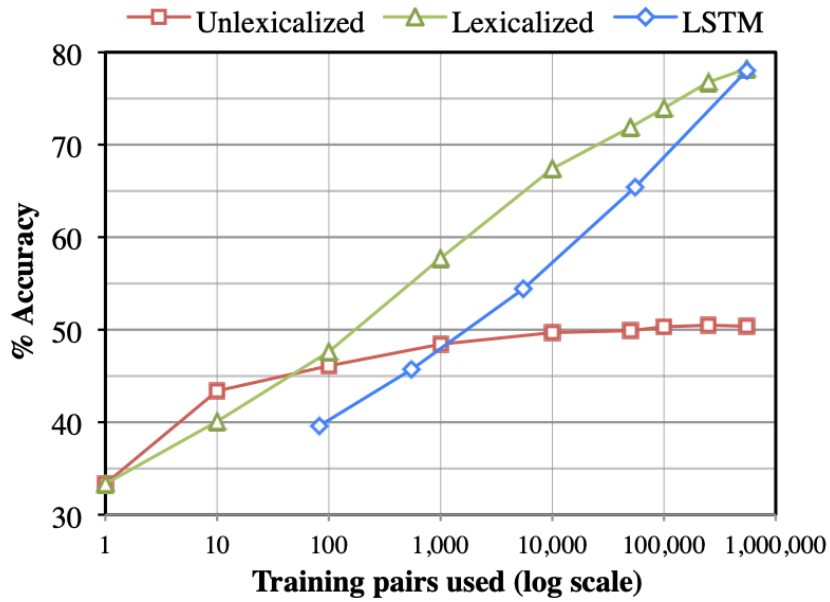


3-class results

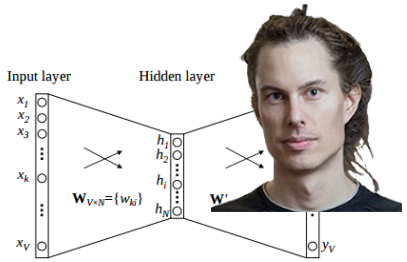
Model	SNLI Train	SNLI Test	SICK Train	SICK Test
Most frequent label	33.4	34.2	56.4	56.7
Classifier	99.7	**78.2**	90.4	**77.8**
no bigrams	93.1	71.6	88.1	77.0
no unigrams/bigrams	49.5	50.4	69.9	69.6
Neural networks				
LSTM RNN	84.8	77.6	–	–
Simple RNN	73.1	72.2	–	–
Sum-of-words	79.3	75.3	–	–

Impact of Data Scaling

- While the lexicalised methods perform well, they are reaching the limits of what they can convey
- Neural methods continue to improve with additional training data
- The paper calls for more data and improved architectures for the neural models.



This paper quite accurately predicted the future of NLP:



2012[2] 2013[3] 2015 2017[4] 2022

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. Commun.
[3] Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations.
[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems