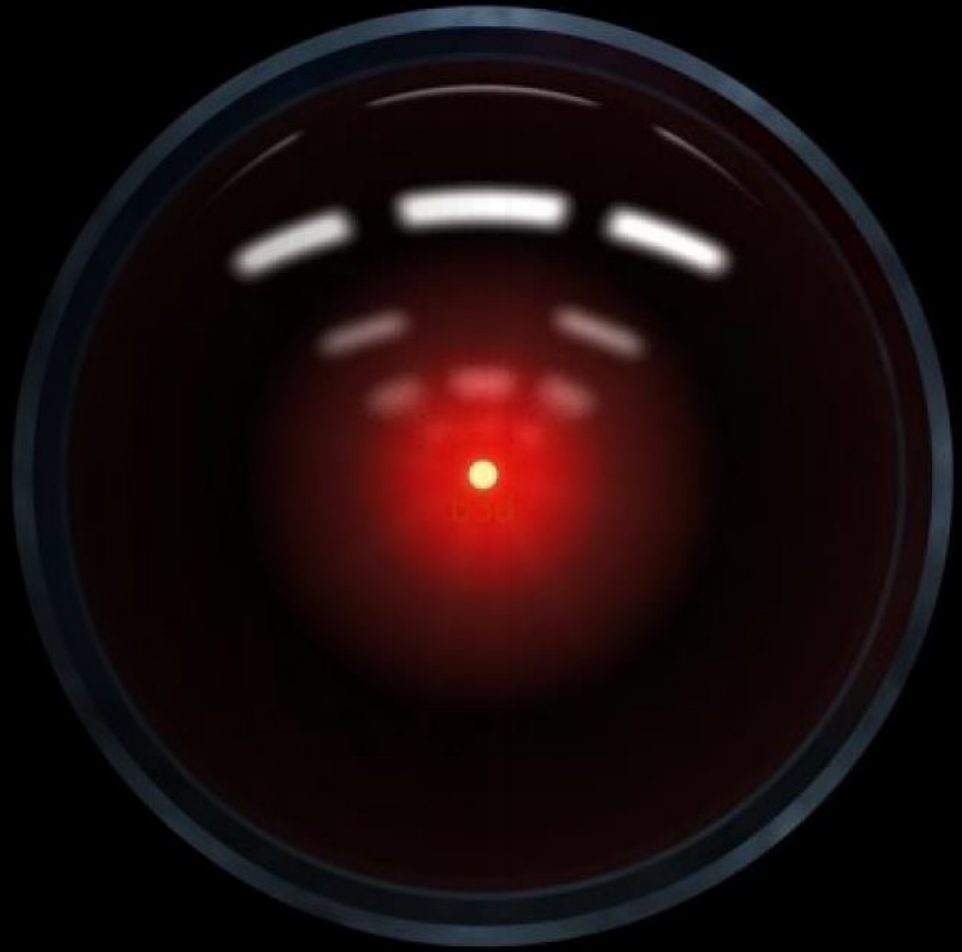
UNIVERSITY OF CAMBRIDGE

A large annotated corpus for learning natural language inference (Bowman et al., 2015)



Presentation for L101: Machine Learning for Language Processing

Raymond Wang

25 October 2022

NLI

p (premise)	h (hypothesis)	?
Exhausted looking firemen are walking.	Firemen are walking.	Entailment
A man walking proudly down the street.	The man is part of the gay pride parade.	Neutral
Two ladies are reading through binders.	The girls are watching a movie.	Contradiction

SNLI (Bowman et al., 2015)

Existing datasets

- Stanford Natural Language Inference Corpus
- 570k pairs of sentences
- Written by humans, labelled by humans

Corpus	Size	Natural	Validated
FraCaS	300	-	Yes
RTE	7k	Yes	Yes
SICK	10k	Yes	Yes
SNLI	570k	Yes	Yes
DG	728k	-	
Levy	1,500k		
PPDB	100,000k	-	

Bowman, Samuel, EMNLP, 2015

Coreference

Bowman et al., 2015

Process

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.*

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

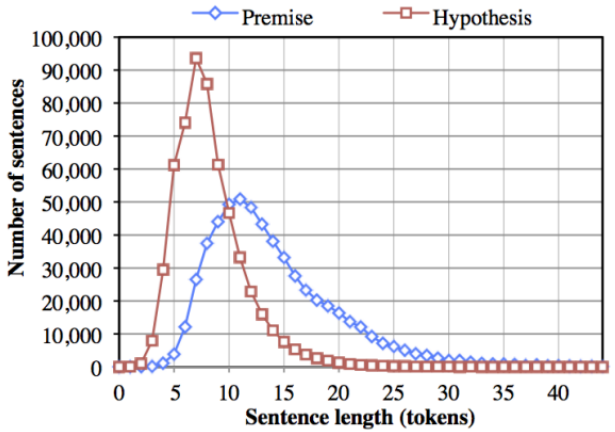


Figure 2: The distribution of sentence length.

Bowman et al., 2015

Validation

Around 10% of data were validated by 4 more annotators

Rate of agreement is extremely high -> corpus is sufficiently high quality to pose a challenging but realistic task

General:	
Validated pairs	56,951
Pairs w/ unanimous gold label	58.3%
Individual annotator label agreement:	
Individual label = gold label	89.0%
Individual label = author's label	85.8%
Gold label/author's label agreement:	
Gold label = author's label	91.2%
Gold label ≠ author's label	6.8%
No gold label (no 3 labels match)	2.0%
Fleiss κ:	
contradiction	0.77
entailment	0.72
neutral	0.60
Overall	0.70

Bowman et al., 2015

Model Results

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

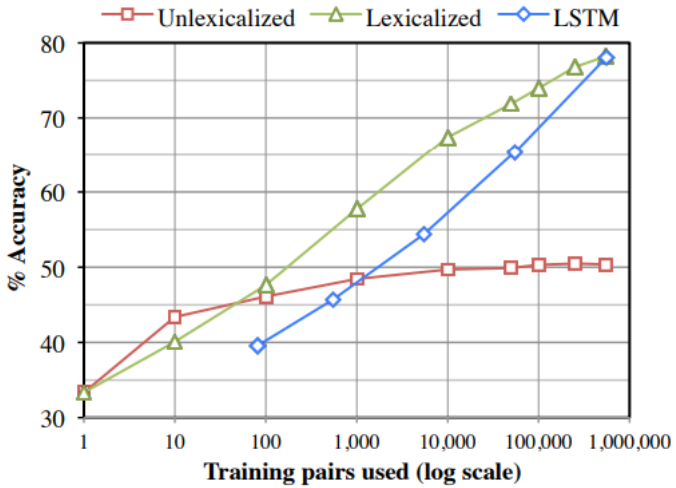
Table 5: 3-class accuracy, training on either our data or SICK, including models lacking cross-bigram features (Feature 6), and lacking all lexical features (Features 4–6). We report results both on the test set and the training set to judge overfitting.

Lexicalised classifier

Training sets	Train	Test
Our data only	42.0	46.7
SICK only	100.0	71.3
Our data and SICK (transfer)	99.9	**80.8**

Table 7: LSTM 3-class accuracy on the SICK train and test sets under three training regimes.

Sentence embeddings



Bowman et al., 2015

Issues

- Captions
- Short sentences
- Lexicalised and unigrams
- Aggregate performance into accuracy
- No world knowledge; word embedding and senses
- h shares focus with p
- Validation

Validation?

p	h	? (e, n, c)
A young woman with long orange hair is sitting on a city bench.	A young woman is sitting on a bench in the park.	Undecidable (2, 2, 1)
Two elderly men having a conversation, snow covered grass in the background.	The men are drinking coffee and having some cookies.	Contradiction (0, 2, 3)
two people working in water next to field	Two people are planting rice.	Neutral (0, 3, 2)
A man in red stands with his child at the beach.	A man wearing red standing with his child at a beach overlooking the ocean.	Neutral (2, 3, 0)
Man wearing black t-shirt sitting at a computer desk.	The man is working on the computer.	Entailment (3, 2, 0)

SNLI (Bowman et al., 2015)

Development

- MNLI (Williams et al., 2018)
- Transformer based models (see BERT Devlin et al., 2018; RoBERTa Liu et al., 2019, etc.)
- Evaluation metrics (see GLUE Wang et al., 2019)

References

- Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. doi:10.48550/ARXIV.1508.05326
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. doi:10.48550/ARXIV.1810.04805
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. doi:10.48550/ARXIV.1907.11692
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. doi:10.48550/ARXIV.1804.07461
- Williams, A., Nangia, N., & Bowman, S. R. (2017). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. doi:10.48550/ARXIV.1704.05426