Inference as benchmark

Inference

- Language model and embedding evaluation is a difficult task
- How can we measure our progress?
- How can we measure that the embeddings reflect an “understanding” of the language

Stanford Natural Language Inference (SNLI) corpus

A large annotated corpus for learning natural language inference

Samuel R. Bowman[*†]
sbowman@stanford.edu

Gabor Angeli[†‡]
angeli@stanford.edu

Christopher Potts[*]
cgpotts@stanford.edu

Christopher D. Manning[*†‡]
manning@stanford.edu

[*]Stanford Linguistics [†]Stanford NLP Group [‡]Stanford Computer Science

https://nlp.stanford.edu/pubs/snli_paper.pdf

SNLI

Neutral A person on a horse jumps over a broken down airplane.

A person is training his horse for a competition.

Contradiction A person on a horse jumps over a broken down airplane.

A person is at a diner, ordering an omelette.

Entailment A person on a horse jumps over a broken down airplane.

A person is outdoors, on a horse.

SNLI

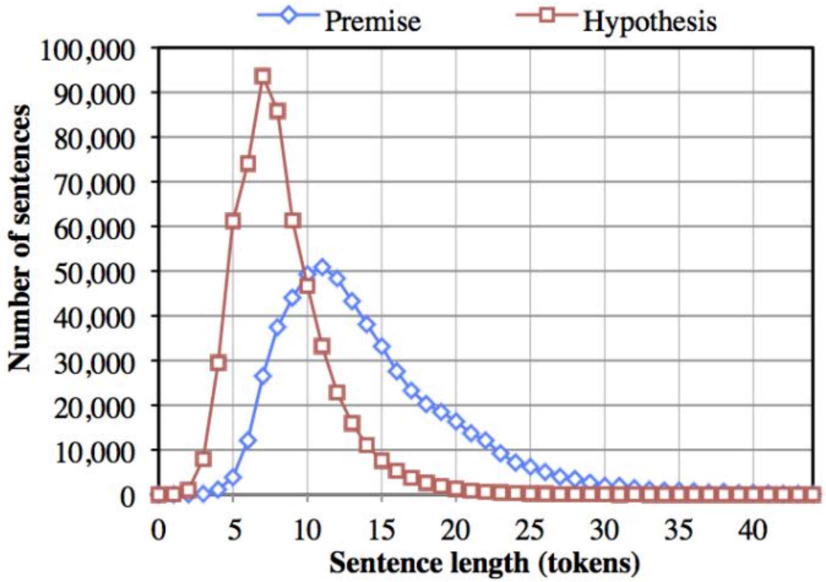

Figure 2: The distribution of sentence length.

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

Table 2: Key statistics for the raw sentence pairs in SNLI. Since the two halves of each pair were collected separately, we report some statistics for both.

MultiNLI + XNLI

MultiNLI

- https://www.nyu.edu/projects/bowman/multinli/
- Modelled after SNLI, but more domains

XNLI

- http://www.nyu.edu/projects/bowman/xnli/
- Multilingual addition to MultiNLI

MultiNLI

Genre	#Examples Train	Dev.	Test	#Wds. Prem.	'S' parses Prem.	Hyp.	Agrmt.	Model Acc. ESIM	CBOW
SNLI	*550,152*	*10,000*	*10,000*	*14.1*	*74%*	*88%*	*89.0%*	*86.7%*	*80.6 %*
FICTION	77,348	2,000	2,000	14.4	94%	97%	89.4%	73.0%	67.5%
GOVERNMENT	77,350	2,000	2,000	24.4	90%	97%	87.4%	74.8%	67.5%
SLATE	77,306	2,000	2,000	21.4	94%	98%	87.1%	67.9%	60.6%
TELEPHONE	83,348	2,000	2,000	25.9	71%	97%	88.3%	72.2%	63.7%
TRAVEL	77,350	2,000	2,000	24.9	97%	98%	89.9%	73.7%	64.6%
9/11	0	2,000	2,000	20.6	98%	99%	90.1%	71.9%	63.2%
FACE-TO-FACE	0	2,000	2,000	18.1	91%	96%	89.5%	71.2%	66.3%
LETTERS	0	2,000	2,000	20.0	95%	98%	90.1%	74.7%	68.3%
OUP	0	2,000	2,000	25.7	96%	98%	88.1%	71.7%	62.8%
VERBATIM	0	2,000	2,000	28.3	93%	97%	87.3%	71.9%	62.7%
MultiNLI Overall	**392,702**	**20,000**	**20,000**	**22.3**	**91%**	**98%**	**88.7%**	**72.2%**	**64.7%**

Table 3: Key statistics for the corpus by genre. The first five genres represent the *matched* section of the development and test sets, and the remaining five represent the *mismatched* section. The first three statistics provide the number of examples in each genre. *#Wds. Prem.* is the mean token count among premise sentences. *'S' parses* is the percentage of sentences for which the Stanford Parser produced a parse rooted with an 'S' (sentence) node. *Agrmt.* is the percent of individual labels that match the gold label in validated examples. *Model Acc.* gives the test accuracy for ESIM and CBOW models (trained on either SNLI or MultiNLI), as described in Section 3.

XNLI

- ❖ Additional development and test sets for MultiNLI
- ❖ 750 pairs x 10 genres x 15 languages = 112,500 annotated pairs
- ❖ English, French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili and Urdu
- ❖ XLM sota (Lample & Conneau 2019) as of Jan. 2019
- ❖ One of the evaluation datasets used in the *cross-lingual transferability of monolingual representations* paper (Artetxe et al.,

Summary

SNLI (S for Stanford)	550k training + 10k dev + 10k test = 570k (pairs)
MNLI (M for multi-genre)	390k training + 20k dev + 20k test = 430k (pairs)
XNLI (X for cross-lingual)	750 pairs x 10 genres x 15 languages = 112,500 annotated pairs
ANLI (A for adversarial)	163k training + 2k dev + 2k test = 167k (pairs)

Annotation artifacts

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are **at least** three **people** on a loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family.**
Contradiction	A woman is **not** taking money for any of her sticks.

Table 1: An instance from SNLI that illustrates the artifacts that arise from the annotation protocol. A common strategy for generating entailed hypotheses is to remove gender or number information. Neutral hypotheses are often constructed by adding a purpose clause. Negations are often introduced to generate contradictions.

Heuristic Analysis for NLI Systems (HANS)

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

Table 1: The heuristics targeted by the HANS dataset, along with examples of incorrect entailment predictions that these heuristics would lead to.

HANS

- ❖ Statistical NLI models may adopt three fallible syntactic heuristics
  - ➢ the lexical overlap heuristic
  - ➢ the subsequence heuristic
  - ➢ the constituent heuristic
- ❖ Models trained on MNLI, including BERT, a state-of-the-art model, perform very poorly on HANS, suggesting that they have indeed adopted these heuristics
- ❖ Augmenting a model’s training set with HANS helps

GLUE and SuperGLUE

- Multi-task benchmark for development and testing of deep language models
- https://gluebenchmark.com/
- Tasks section lists the tasks forming the benchmarks

Recap

- Inference-type data is a good performance benchmark in theory
- Producing good datasets of this kind is surprisingly difficult
- The models can learn based on surface cues rather than actual meaning
- Composite scores like GLUE measuring across a number of tasks are the present standard
- Ongoing research, fast-moving target