A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference

Adina Williams, Nikita Nangia, Samuel R. Bowman

Reporter. LiuYang

ADINA WILLIAMS

- PhD candidate in Linguistics at New York University
- semantics, syntax, and their interface, with neurolinguistics, computational linguistics and formal linguistic theory
- NYU Neuroscience of Language Lab (NELLab; Liina Pylkkänen), and with the Machine Learning for Language (ML² Group; Sam Bowman).

Examples

Met my first girlfriend that way.	FACE-TO-FACE **contradiction** C C N C	I didn't meet my first girlfriend until later.
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** N N N N	The 8 million dollars for emergency housing was still not enough to solve the problem.
Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	LETTERS **neutral** N N N N	All of the children love working in their gardens.
At 8:34, the Boston Center controller received a third transmission from American 11	9/11 **entailment** E E E E	The Boston Center controller got a third transmission from American 11.
I am a lacto-vegetarian.	SLATE **neutral** N N E N	I enjoy eating cheese too much to abstain from dairy.
someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny	TELEPHONE **contradiction** C C C C	No one noticed and it wasn't funny at all.

Table 1: Randomly chosen examples from the development set of our new corpus, shown with their genre labels, their selected gold labels, and the validation labels (abbreviated E, N, C) assigned by individual annotators.

Motivation

- research on the core problems of NLU
- domain adaptation and cross-domain transfer learning
  - General purpose feature extractors (In other ML task not NLU)
  - Multi-NLI :
    - 10 different genres of written and spoken English
    - 5 genres in training data
    - 10 genres in dev/test data

Data Collection

- maximally diverse and roughly represent the full range of American English
- nine sources from Open American National Corpus (OANC) balancing across genres
- minimal preprocessing
- SNLI can be appended and treated as an unusually large additional CAPTIONS genre

- In-person conversations (FACE-TO-FACE);
- government websites (GOVERNMENT);
- letters from Philanthropic Fundraising (LETTERS);
- Terrorist Attacks(9/11);
- Oxford University Press (OUP)
- Slate Magazine (SLATE)
- …

Data Collection

- selecting a premise sentence from a preexisting text source and asking a human annotator to compose a novel sentence to pair with it as a hypothesis
- Hybrid (gethybrid.io) 387 annotators
- additional round of annotation (Each pair is relabeled by four workers)
- label one percent on the validation examples and offer a $1 if match

- Write one sentence that is definitely correct about the situation or event in the line.
- Writeone sentence that might be correct about the situation or event in the line.
- Write one sentence that is definitely incorrect about the situation or event in the line.

Data Collection

- MultiNLI are about as reliable as those included in SNLI, despite MultiNLI's more diverse text contents.

Statistic	SNLI	MultiNLI
Pairs w/ unanimous gold label	58.3%	58.2%
Individual label = gold label	89.0%	88.7%
Individual label = author's label	85.8%	85.2%
Gold label = author's label	91.2%	92.6%
Gold label ≠ author's label	6.8%	5.6%
No gold label (no 3 labels match)	2.0%	1.8%

Table 2: Key validation statistics for SNLI (copied from Bowman et al., 2015) and MultiNLI.

Corpus Result

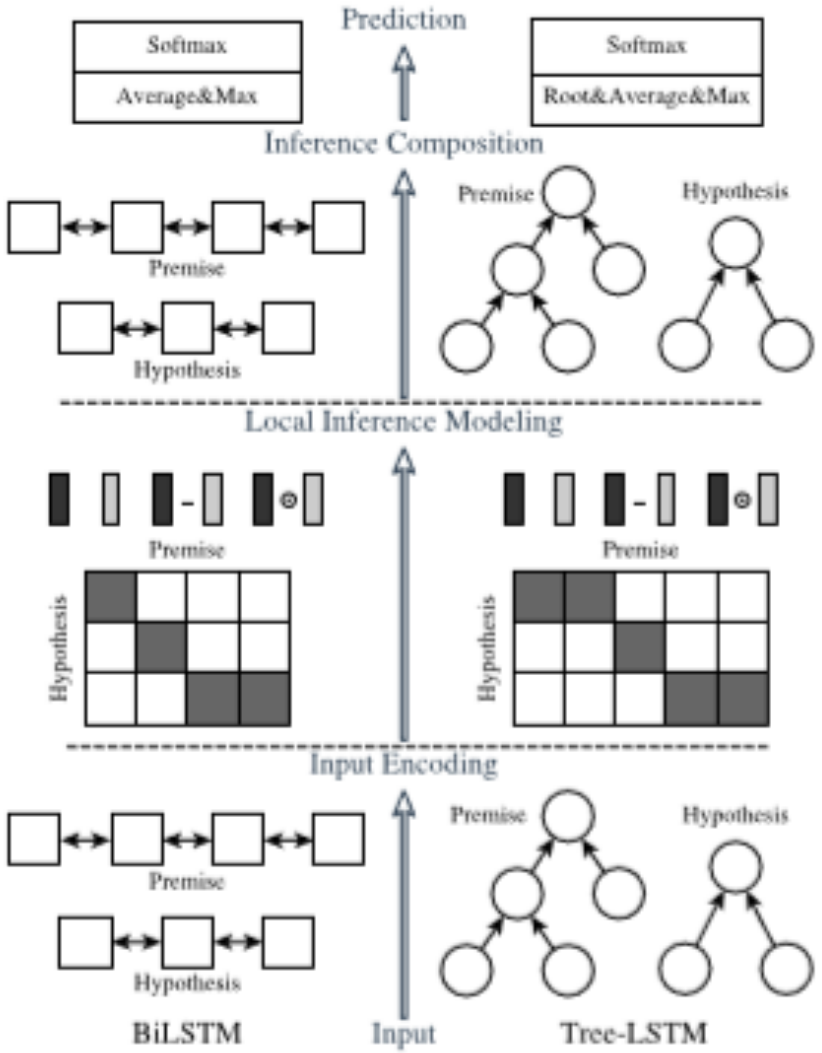
- freely available at nyu.edu/projects/bowman/multinli
- modified and redistributed
- released under the OANC's license

Corpus Result

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%**

Baseline

- Model
  - simple continuous bag of words (CBOW)
  - averaging the states of a bidirectional LSTM RNN (BiLSTM)
  - Chen et al.'s Enhanced Sequential Inference Model (ESIM)



Baseline Results

- Model Input
  - Full Multi-NLI + 15% SNLI (randomly)
  - Word Embedding (300d GloVe) => representations for each sentence
  - premise and hypothesis, their difference, and their element-wise product (first 2 model)

Train	Model	SNLI	MNLI Match.	MNLI Mis.
	Most freq.	34.3	36.5	35.6
SNLI	CBOW	80.6	-	-
	BiLSTM	81.5	-	-
	ESIM	**86.7**	-	-
MNLI	CBOW	51.5	64.8	64.5
	BiLSTM	50.8	66.9	66.9
	ESIM	60.7	**72.3**	**72.1**
MNLI+ SNLI	CBOW	74.7	65.2	64.6
	BiLSTM	74.0	67.5	67.1
	ESIM	79.7	**72.4**	**71.9**

Table 4: Test set accuracies (%) for all models; *Match.* represents test set performance on the MultiNLI genres that are also represented in the training set, *Mis.* represents test set performance on the remaining ones; *Most freq.* is a trivial 'most frequent class' baseline.

Discussion and Analysis

- SNLI - dependent on the concreteness of image descriptions
- Multi-NLI - design prompts for abstract genres

- *a boat sank in the Pacific Ocean*
- *a boat sank in the Atlantic Ocean*

Discussion and Analysis

- Difficulty
  - increase diversity of linguistic phenomena
  - longer average sentence length
- No help when add 15% SNLI

Train	Model	SNLI	MNLI Match.	MNLI Mis.
	Most freq.	34.3	36.5	35.6
SNLI	CBOW	80.6	-	-
	BiLSTM	81.5	-	-
	ESIM	**86.7**	-	-
MNLI	CBOW	51.5	64.8	64.5
	BiLSTM	50.8	66.9	66.9
	ESIM	60.7	**72.3**	**72.1**
MNLI+ SNLI	CBOW	74.7	65.2	64.6
	BiLSTM	74.0	67.5	67.1
	ESIM	79.7	**72.4**	**71.9**

Table 4: Test set accuracies (%) for all models; *Match.* represents test set performance on the MultiNLI genres that are also represented in the training set, *Mis.* represents test set performance on the remaining ones; *Most freq.* is a trivial 'most frequent class' baseline.

Discussion and Analysis

- Penn Treebank (PTB) part-of-speech tag set automatically isolate sentences containing a range of easily-identified phenomena like comparatives
- contain negation are more likely to be labeled CONTRADICTION
- long sentences are more likely to be labeled ENTAILMENT.
- on discourse markers, such as despite and however, losing roughly 2 to 3 points
- ESIM model performs on sentences with greater than 20 words

Tag	Dev. Freq. SNLI	Dev. Freq. MultiNLI	Diff.	Most Frequent Label	Most Frequent Label %
Entire Corpus	100	100	0	entailment	∼35
Pronouns (PTB)	34	68	34	entailment	34
Quantifiers	33	63	30	contradiction	36
Modals (PTB)	<1	28	28	entailment	35
Negation (PTB)	5	31	26	contradiction	**48**
WH terms (PTB)	5	30	25	entailment	35
Belief Verbs	<1	19	18	entailment	34
Time Terms	19	36	17	neutral	35
Discourse Mark.	<1	14	14	neutral	34
Presup. Triggers	8	22	14	neutral	34
Compr./Supr.(PTB)	3	17	14	neutral	39
Conditionals	4	15	11	neutral	35
Tense Match (PTB)	62	69	7	entailment	37
Interjections (PTB)	<1	5	5	entailment	36
>20 words	<1	5	5	entailment	**42**

Conclusion

- NLI makes it easy to judge NLU
- A new dataset that offers dramatically greater linguistic difficulty and diversity
- Multi-NLI has a lot of headroom remaining for future work

Thanks and Q&A.