Evaluating Recent Progress Toward General-Purpose Language Understanding Models

Sam Bowman
@sleepinyourhat
The Goal

To develop a general-purpose neural network encoder for text which makes it possible to solve any new language understanding task using only enough training data to define the possible outputs.
The Goal

To develop a neural network model that already understands English when it starts learning a new task.
The Technique: Muppets

Large-scale pretrained language models like ELMo, GPT, BERT, XLNet, RoBERTa, and T5 have offered a recent surge of progress toward this goal.
This Talk

• The GLUE language understanding benchmark
  Wang et al. '19a

• Recent progress and the updated SuperGLUE benchmark
  Nangia & Bowman '19, Wang et al. '19b

• A few things we've learned about modern models
  Tenney et al. '19, Warstadt et al. '19

• What's next for evaluation?
  Idle speculation '19
GLUE: What is it?
The General Language Understanding Evaluation (GLUE):

An open-ended competition and evaluation platform for general-purpose sentence encoders.
Why GLUE?

Increasingly common for researchers outside NLP to evaluate new techniques on language understanding tasks.

• We can learn a lot this way...

• ...if these researchers evaluate on significant open problems...

• ...which doesn't always happen.

Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19
Why GLUE?

GLUE for non-NLP-specialist researchers:

• We provide tasks, metrics, baselines, and code that represent open problems of interest to researchers in NLU.

• We don't enforce any particular experimental design —that's up to the (expert) users.
Nine English-language sentence understanding tasks based on existing data:

- Unsolved
- Varied training data volume
- Varied language style/genre
Simple task APIs:

- Only sentence or sentence pair inputs.
- Only classification or regression outputs.
- *No generation or structured prediction.*
Simple leaderboard API: Upload predictions for a test set (like Kaggle/SemEval)

- Usable with any software infrastructure.
- Usable with any kind of method/model!
- Allows us to limit use of the test sets.
## GLUE: The Main Tasks

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoLA</td>
<td>8.5k</td>
<td>1k</td>
<td>1k</td>
<td>acceptability</td>
<td>Matthews corr.</td>
<td>misc.</td>
</tr>
<tr>
<td>SST-2</td>
<td>67k</td>
<td>872</td>
<td>1.8k</td>
<td>sentiment</td>
<td>acc.</td>
<td>movie reviews</td>
</tr>
</tbody>
</table>

**Single-Sentence Tasks**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRPC</td>
<td>3.7k</td>
<td>408</td>
<td>1.7k</td>
<td>paraphrase</td>
<td>acc./F1</td>
<td>news</td>
</tr>
<tr>
<td>STS-B</td>
<td>7k</td>
<td>1.5k</td>
<td>1.4k</td>
<td>sentence similarity</td>
<td>Pearson/Spearman corr.</td>
<td>misc.</td>
</tr>
<tr>
<td>QQP</td>
<td>364k</td>
<td>40k</td>
<td>391k</td>
<td>paraphrase</td>
<td>acc./F1</td>
<td>social QA questions</td>
</tr>
</tbody>
</table>

**Similarity and Paraphrase Tasks**

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNLI</td>
<td>393k</td>
<td>20k</td>
<td>20k</td>
<td>NLI</td>
<td>matched acc./mismatched acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>QNLI</td>
<td>108k</td>
<td>5.7k</td>
<td>5.7k</td>
<td>QA/NLI</td>
<td>acc.</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>RTE</td>
<td>2.5k</td>
<td>276</td>
<td>3k</td>
<td>NLI</td>
<td>acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>WNLI</td>
<td>634</td>
<td>71</td>
<td>146</td>
<td>coreference/NLI</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>

**Inference Tasks**

Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19
# GLUE: The Main Tasks

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoLA</td>
<td>8.5k</td>
<td>1k</td>
<td>1k</td>
<td>acceptability</td>
<td>Matthews corr.</td>
<td>misc.</td>
</tr>
<tr>
<td>SST-2</td>
<td>67k</td>
<td>872</td>
<td>1.8k</td>
<td>sentiment</td>
<td>acc.</td>
<td>movie reviews</td>
</tr>
<tr>
<td>MRPC</td>
<td>3.7k</td>
<td>408</td>
<td>1.7k</td>
<td>paraphrase</td>
<td>acc./F1</td>
<td>news</td>
</tr>
<tr>
<td>STS-B</td>
<td>7k</td>
<td>1.5k</td>
<td>1.4k</td>
<td>sentence similarity</td>
<td>Pearson/Spearman corr.</td>
<td>misc.</td>
</tr>
<tr>
<td>QQP</td>
<td>364k</td>
<td>40k</td>
<td>391k</td>
<td>paraphrase</td>
<td>acc./F1</td>
<td>social QA questions</td>
</tr>
<tr>
<td>MNLI</td>
<td>393k</td>
<td>20k</td>
<td>20k</td>
<td>NLI</td>
<td>matched acc./mismatched acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>QNLI</td>
<td>108k</td>
<td>5.7k</td>
<td>5.7k</td>
<td>QA/NLI</td>
<td>acc.</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>RTE</td>
<td>2.5k</td>
<td>276</td>
<td>3k</td>
<td>NLI</td>
<td>acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>WNLI</td>
<td>634</td>
<td>71</td>
<td>146</td>
<td>coreference/NLI</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>
## GLUE: The Main Tasks

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Single-Sentence Tasks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoLA</td>
<td>8.5k</td>
<td>1k</td>
<td>1k</td>
<td>acceptability</td>
<td>Matthews corr.</td>
<td>misc.</td>
</tr>
<tr>
<td>SST-2</td>
<td>67k</td>
<td>872</td>
<td>1.8k</td>
<td>sentiment</td>
<td>acc.</td>
<td>movie reviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Similarity and Paraphrase Tasks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRPC</td>
<td>3.7k</td>
<td>408</td>
<td>1.7k</td>
<td>paraphrase</td>
<td>acc./F1</td>
<td>news</td>
</tr>
<tr>
<td>STS-B</td>
<td>7k</td>
<td>1.5k</td>
<td>1.4k</td>
<td>sentence similarity</td>
<td>Pearson/Spearman corr.</td>
<td>misc.</td>
</tr>
<tr>
<td>QQP</td>
<td>364k</td>
<td>40k</td>
<td>391k</td>
<td>paraphrase</td>
<td>acc./F1</td>
<td>social QA questions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inference Tasks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MNLI</td>
<td>393k</td>
<td>20k</td>
<td>20k</td>
<td>NLI</td>
<td>matched acc./mismatched acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>QNLI</td>
<td>108k</td>
<td>5.7k</td>
<td>5.7k</td>
<td>QA/NLI</td>
<td>acc.</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>RTE</td>
<td>2.5k</td>
<td>276</td>
<td>3k</td>
<td>NLI</td>
<td>acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>WNL1</td>
<td>634</td>
<td>71</td>
<td>146</td>
<td>coreference/NLI</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>

Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19
## GLUE: The Main Tasks

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoLA</td>
<td>8.5k</td>
<td>1k</td>
<td>1k</td>
<td>acceptability</td>
<td>Matthews corr.</td>
<td>misc.</td>
</tr>
<tr>
<td>SST-2</td>
<td>67k</td>
<td>872</td>
<td>1.8k</td>
<td>sentiment</td>
<td>acc.</td>
<td>movie reviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRPC</td>
<td>3.7k</td>
<td>408</td>
<td>1.7k</td>
<td>paraphrase</td>
<td>acc./F1</td>
<td>news</td>
</tr>
<tr>
<td>STS-B</td>
<td>7k</td>
<td>1.5k</td>
<td>1.4k</td>
<td>sentence similarity</td>
<td>Pearson/Spearman corr.</td>
<td>misc.</td>
</tr>
<tr>
<td>QQP</td>
<td>364k</td>
<td>40k</td>
<td>391k</td>
<td>paraphrase</td>
<td>acc./F1</td>
<td>social QA questions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MNLI</td>
<td>393k</td>
<td>20k</td>
<td>20k</td>
<td>NLI</td>
<td>matched acc./mismatched acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>QNLI</td>
<td>108k</td>
<td>5.7k</td>
<td>5.7k</td>
<td>QA/NLI</td>
<td>acc.</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>RTE</td>
<td>2.5k</td>
<td>276</td>
<td>3k</td>
<td>NLI</td>
<td>acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>WNLNI</td>
<td>634k</td>
<td>71</td>
<td>146</td>
<td>coreference/NLI</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>
## GLUE: The Main Tasks

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoLA</td>
<td>8.5k</td>
<td>1k</td>
<td>1k</td>
<td>acceptability</td>
<td>Matthews corr.</td>
<td>misc.</td>
</tr>
<tr>
<td>SST-2</td>
<td>67k</td>
<td>872</td>
<td>1.8k</td>
<td>sentiment</td>
<td>acc.</td>
<td>movie reviews</td>
</tr>
<tr>
<td>MRPC</td>
<td>3.7k</td>
<td>408</td>
<td>1.7k</td>
<td>paraphrase</td>
<td>acc./F1</td>
<td>news</td>
</tr>
<tr>
<td>STS-B</td>
<td>7k</td>
<td>1.5k</td>
<td>1.4k</td>
<td>sentence similarity</td>
<td>Pearson/Spearman corr.</td>
<td>misc.</td>
</tr>
<tr>
<td>QQP</td>
<td>364k</td>
<td>40k</td>
<td>391k</td>
<td>paraphrase</td>
<td>acc./F1</td>
<td>social QA questions</td>
</tr>
<tr>
<td>MNLI</td>
<td>393k</td>
<td>20k</td>
<td>20k</td>
<td>NLI</td>
<td>matched acc./mismatched acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>QNLI</td>
<td>108k</td>
<td>5.7k</td>
<td>5.7k</td>
<td>QA/NLI</td>
<td>acc.</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>RTE</td>
<td>2.5k</td>
<td>276</td>
<td>3k</td>
<td>NLI</td>
<td>acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>WNLI</td>
<td>634</td>
<td>71</td>
<td>146</td>
<td>coreference/NLI</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>

Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19
The Corpus of Linguistic Acceptability (CoLA)

Warstadt et al. '18

- Binary classification: Is some string of words a possible English sentence.
- Data of this form is a major source of evidence in linguistic theory. Sentences derived from books and articles on morphology, syntax, and semantics.

* Who do you think that will question Seamus first?
✓ The gardener planted roses in the garden.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoLA</td>
<td>8.5k</td>
<td>1k</td>
<td>1k</td>
<td>acceptability</td>
<td>Matthews corr.</td>
<td>misc.</td>
</tr>
<tr>
<td>SST-2</td>
<td>67k</td>
<td>872</td>
<td>1.8k</td>
<td>sentiment</td>
<td>acc.</td>
<td>movie reviews</td>
</tr>
</tbody>
</table>

Wang, Singh, Michael, Hill, Levy & Bowman ICLR '19
The Recognizing Textual Entailment Challenge

Dagan et al. '06 et seq.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoLA SST-2</td>
<td>• Binary classification over sentence pairs: Does the first sentence entail the second?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Drawn from several of the RTE annual competitions.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRPC STS-B QQP</td>
<td>Text: Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hypothesis: Christopher Reeve had an accident.</td>
<td></td>
<td></td>
<td>no-entailment</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Corup</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNLI</td>
<td>393k</td>
<td>20k</td>
<td>20k</td>
<td>NLI</td>
<td>matched acc./mismatched acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>QNLI</td>
<td>108k</td>
<td>5.7k</td>
<td>5.7k</td>
<td>QA/NLI</td>
<td>acc.</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>RTE</td>
<td>2.5k</td>
<td>276</td>
<td>3k</td>
<td>NLI</td>
<td>acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>WNLNI</td>
<td>634</td>
<td>71</td>
<td>146</td>
<td>coreference/NLI</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>
The Winograd Schema Challenge
NLI format, based on Levesque et al., 2011

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoLA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SST-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRPC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STS-B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QQP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MNLI</td>
<td>393k</td>
<td>20k</td>
<td>20k</td>
<td>NLI</td>
<td>matched acc./mismatched acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>QNLI</td>
<td>108k</td>
<td>5.7k</td>
<td>5.7k</td>
<td>QA/NLI</td>
<td>acc.</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>RTE</td>
<td>2.5k</td>
<td>276</td>
<td>3k</td>
<td>NLI</td>
<td>acc.</td>
<td>misc.</td>
</tr>
<tr>
<td>WNLI</td>
<td>634</td>
<td>71</td>
<td>146</td>
<td>coreference/NLI</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>

- Binary classification for expert-constructed pairs of sentences: What does the pronoun refer to?
- Manually constructed to foil superficial statistical cues.
- Private evaluation data used only in GLUE.

P: Jane gave Joan candy because she was hungry.
H: Joan was hungry.

entailment
GLUE: What methods work?
GLUE Score: Highlights

95
85
75
65
55

GloVe BoW
Single-Task Models
Sentence-to-Vector
ELMo
OpenAI GPT
BERT Large
Human Crowdworkers
MT-DNN
RoBERTa
ALBERT
T5

gluebenchmark.com
GLUE Score: Highlights

A strong baseline without access to word order.
The preexisting standard practice: One model per task, trained from scratch.
The state of the art from earlier pretraining work: A fixed sentence-to-vector encoder.
Concurrent work, and the first major success with language model pretraining.
The first transfer model to use a *Transformer* architecture, plus fine-tuning for target tasks.
GLUE Score: Highlights

Shift from left-to-right language modeling to omnidirectional masked language modeling.
How much headroom does GLUE have left?

- To compute a conservative estimate for each task:
  - Train crowdworkers.
How much headroom does GLUE have left?

• To compute a conservative estimate for each task:
  • Train crowdworkers.
  • Get multiple crowdworker labels for each example, take a majority vote.
GLUE Score: Highlights

- GloVe BoW
- Single-Task Models
- Sentence-to-Vector
- ELMo
- OpenAI GPT
- BERT Large
- Human Crowdworkers
- MT-DNN
- RoBERTa
- ALBERT
- T5
GLUE Score: Highlights

Sharing information across the nine target tasks.
GLUE Score: Highlights

- Longer training, more data.
GLUE Score: Highlights

More efficient parameterization, bigger model.
Even bigger model, joint training across lots of labeled data tasks.
We rebuilt GLUE from scratch...

• ...starting with an open call for dataset proposals

• ...yielding 30–40 candidates

• ...which we filtered using human evaluation and BERT-base baselines

• ...and a final set of eight tasks

• ...following a slightly expanded set of task APIs.
## SuperGLUE: The Main Tasks

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Text Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB</td>
<td>250</td>
<td>57</td>
<td>250</td>
<td>NLI</td>
<td>acc./F1</td>
<td>various</td>
</tr>
<tr>
<td>COPA</td>
<td>400</td>
<td>100</td>
<td>500</td>
<td>QA</td>
<td>acc.</td>
<td>blogs, photography encyclopedia</td>
</tr>
<tr>
<td>MultiRC</td>
<td>5100</td>
<td>953</td>
<td>1800</td>
<td>QA</td>
<td>F1_a/EM</td>
<td>various</td>
</tr>
<tr>
<td>ReCoRD</td>
<td>101k</td>
<td>10k</td>
<td>10k</td>
<td>QA</td>
<td>F1/EM</td>
<td>news (CNN, Daily Mail)</td>
</tr>
<tr>
<td>RTE</td>
<td>2500</td>
<td>278</td>
<td>300</td>
<td>NLI</td>
<td>acc.</td>
<td>news, Wikipedia</td>
</tr>
<tr>
<td>WiC</td>
<td>6000</td>
<td>638</td>
<td>1400</td>
<td>WSD</td>
<td>acc.</td>
<td>WordNet, VerbNet, Wiktionary</td>
</tr>
<tr>
<td>WSC</td>
<td>554</td>
<td>104</td>
<td>146</td>
<td>coref.</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>

{Wang, Pruksachatkun, Nangia, Singh}, Michael, Hill, Levy & Bowman NeurIPS '19
# SuperGLUE: The Main Tasks

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Text Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB</td>
<td>250</td>
<td>57</td>
<td>250</td>
<td>NLI</td>
<td>acc./F1</td>
<td>various</td>
</tr>
<tr>
<td>COPA</td>
<td>400</td>
<td>100</td>
<td>500</td>
<td>QA</td>
<td>acc.</td>
<td>blogs, photography encyclopedia</td>
</tr>
<tr>
<td>MultiRC</td>
<td>5100</td>
<td>953</td>
<td>1800</td>
<td>QA</td>
<td>F1_a/EM</td>
<td>various</td>
</tr>
<tr>
<td>ReCoRD</td>
<td>101k</td>
<td>10k</td>
<td>10k</td>
<td>QA</td>
<td>F1/EM</td>
<td>news (CNN, Daily Mail)</td>
</tr>
<tr>
<td>RTE</td>
<td>2500</td>
<td>278</td>
<td>300</td>
<td>NLI</td>
<td>acc.</td>
<td>news, Wikipedia</td>
</tr>
<tr>
<td>WiC</td>
<td>6000</td>
<td>638</td>
<td>1400</td>
<td>WSD</td>
<td>acc.</td>
<td>WordNet, VerbNet, Wiktionary</td>
</tr>
<tr>
<td>WSC</td>
<td>554</td>
<td>104</td>
<td>146</td>
<td>coref.</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>
# SuperGLUE: The Main Tasks

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Text Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB</td>
<td>250</td>
<td>57</td>
<td>250</td>
<td>NLI</td>
<td>acc./F1</td>
<td>various</td>
</tr>
<tr>
<td>COPA</td>
<td>400</td>
<td>100</td>
<td>500</td>
<td>QA</td>
<td>acc.</td>
<td>blogs, photography encyclopedia</td>
</tr>
<tr>
<td>MultiRC</td>
<td>5100</td>
<td>953</td>
<td>1800</td>
<td>QA</td>
<td>F1\alpha/EM</td>
<td>various</td>
</tr>
<tr>
<td>ReCoRD</td>
<td>101k</td>
<td>10k</td>
<td>10k</td>
<td>QA</td>
<td>F1/EM</td>
<td>news (CNN, Daily Mail)</td>
</tr>
<tr>
<td>RTE</td>
<td>2500</td>
<td>278</td>
<td>300</td>
<td>NLI</td>
<td>acc.</td>
<td>news, Wikipedia</td>
</tr>
<tr>
<td>WiC</td>
<td>6000</td>
<td>638</td>
<td>1400</td>
<td>WSD</td>
<td>acc.</td>
<td>WordNet, VerbNet, Wiktionary</td>
</tr>
<tr>
<td>WSC</td>
<td>554</td>
<td>104</td>
<td>146</td>
<td>coref.</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>
Three-way NLI classification: Does a speaker utterance entail some embedded clause within that utterance?

Text:
B: And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out.
A: Uh-huh.
B: What do you think, do you think we are, setting a trend?

Hypothesis:
they are setting a trend
no-entailment
• Multiple choice reading comprehension QA over paragraphs.

**Paragraph:** Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week.

**Question:** Did Susan’s sick friend recover?

**Answers:** Yes, she recovered (T), No (F), Yes (T), No, she didn’t recover (F), Yes, she was at Susan’s party (T)
# SuperGLUE: The Main Tasks

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Task</th>
<th>Metrics</th>
<th>Text Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB</td>
<td>250</td>
<td>57</td>
<td>250</td>
<td>NLI</td>
<td>acc./F1</td>
<td>various</td>
</tr>
<tr>
<td>COPA</td>
<td>400</td>
<td>100</td>
<td>500</td>
<td>QA</td>
<td>acc.</td>
<td>blogs, photography encyclopedia</td>
</tr>
<tr>
<td>MultiRC</td>
<td>5100</td>
<td>953</td>
<td>1800</td>
<td>QA</td>
<td>F1_a/EM</td>
<td>various</td>
</tr>
<tr>
<td>ReCoRD</td>
<td>101k</td>
<td>10k</td>
<td>10k</td>
<td>QA</td>
<td>F1/EM</td>
<td>news (CNN, Daily Mail)</td>
</tr>
<tr>
<td>RTE</td>
<td>2500</td>
<td>278</td>
<td>300</td>
<td>NLI</td>
<td>acc.</td>
<td>news, Wikipedia</td>
</tr>
<tr>
<td>WiC</td>
<td>6000</td>
<td>638</td>
<td>1400</td>
<td>WSD</td>
<td>acc.</td>
<td>WordNet, VerbNet, Wiktionary</td>
</tr>
<tr>
<td>WSC</td>
<td>554</td>
<td>104</td>
<td>146</td>
<td>coref.</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>
SuperGLUE Score: Highlights

- GloVe Bag of Words: 45
- BERT: 67
- RoBERTa: 82.5
- T5: 87.5
- Human Crowdworkers: 95

Wang et al. '18

super.gluebenchmark.com
GLUE and SuperGLUE: Limitations

GLUE and SuperGLUE are built only on English data.

- General-purpose pretraining may look quite different in lower-resource languages!
GLUE and SuperGLUE: Limitations

GLUE and SuperGLUE use lots of naturally occurring or crowdsourced data.

- Therefore safe to presume that these datasets contain evidence of social bias (see Rudinger et al., EthNLP '17).

- All else being equal, models that learn and use these biases will do better on these benchmarks.

- In SuperGLUE's WinoGender Schema evaluation (Rudinger et al. ’18), T5 is 10x more like than humans to be confused by irrelevant gender cues.

- Mitigating these biases is a major open problem.
GLUE and SuperGLUE: Non-Limitations

GLUE and SuperGLUE don't test generation or structured prediction.

- These are hand and important problems, but mostly orthogonal to language understanding.
We clearly haven't solved NLU.

SuperGLUE includes a broad-coverage NLI diagnostic:

**Prepositional phrases section**

- *I ate pizza with olives.* *I ate pizza with some friends.*
- *I ate olives.* *I ate some friends.*
- entailment  neutral
We can be pretty sure we haven't solved NLU even for IID evaluations.

- 6-point gap between T5 and humans on Winograd Schemas.
- *In-domain* evaluation for NLI, QA, etc., involves lots of phenomena that we know models aren't great at. Are these differences just drowned in the noise?
Why does BERT* work so well?
What does BERT know?

*Yes, BERT.
What’s inside BERT?

In our work on *Edge Probing* (Tenney et al.), we observe that:

- ELMo and BERT both learn nearly perfect features for POS tagging.

- BERT learns better features than ELMo for parsing.

- ELMo and BERT Base do not learn coreference features, but BERT Large does.
What’s inside BERT?
What’s inside BERT?

In further edge probing studies (Tenney, Das, and Pavlick):

• Lower layers of BERT express features for 'lower level' tasks.

• Higher layers express more abstract/semantic knowledge.
What’s inside BERT?

Structural probes (Hewitt and Manning):

• The geometry of BERT's activation vectors encode some syntactic structure.
What’s inside BERT?

Evaluations on *handbuilt test sets* (Yaghoobzadeh et al.):

- BERT relies on brittle non-syntactic heuristics for tasks like NLI; but BERT Large much less so than BERT Base.
How much can we trust these conclusions?
How much can we trust these conclusions?

• Probing studies (loosely defined) like these are a **common tool** for trying to understand what models like BERT know.

• There are many ways to design such a study, and each bakes in substantial assumptions.

  • Edge probing assumes that if a model *knows* about coreference, then it should be possible to extract that information with a simple MLP model.

• **Do different probing methods give us the same answer?**
Case Study: NPI Licensing

NPI words like *any* or *ever* can only occur in the scope of specific linguistic *licensing environments* like negations or conditionals.

- Well-characterized in the linguistics literature.
- Depends on long-distance dependencies and complex structures, rather than local co-occurrence.

*Does BERT know where NPIs are licensed?*

I see kids who are not [eating *any* cookies].

*I see *any* kids who are not [eating cookies].

{Warstadt, Cao, Grosu, Peng, Blix, Nie, Alsop, Bordia, Liu, Parrish, Wang, Phang, Mohananey, Htut, Jeretič} & Bowman EMNLP ‘19
Case Study: NPI Licensing

NPI words like *any* or *ever* can only occur in the scope of specific linguistic licensing environments like negations or conditionals.

- Well-characterized in the linguistics literature.
- Depends on long-distance dependencies and complex structures, rather than local co-occurrence.

Does BERT know where NPIs are licensed?

Let's ask this as many ways as we can!

-I see kids who are not [eating *any* cookies].

*I see *any* kids who are not [eating cookies].

{Warstadt, Cao, Grosu, Peng, Blix, Nie, Alsop, Bordia, Liu, Parrish, Wang, Phang, Mohananey, Htut, Jeretič} & Bowman
EMNLP ‘19
Case Study: NPI Licensing

Evaluation data: Nine custom NPI test sets isolating different NPI licensors:

*Those boys say that [the doctors ever went to an art gallery.]
*Those boys ever say that [the doctors went to an art gallery.]
Those boys say that [the doctors often went to an art gallery.]
Those boys often say that [the doctors went to an art gallery.]
Let's teach the model to judge acceptability.

*Who do you think that will question Seamus first?
*Usually, any lion is majestic.
The gardener planted roses in the garden.
I wrote Blair a letter, but I tore it up before I sent it.

**Train:**
The CoLA general acceptability corpus

**Test:**
NPI environment test sets

**Metric:**
Matthews Correlation (MCC) for acceptability
Let's teach the model to judge acceptability.

BERT knows a bit about NPIs, but it's not perfect.

* When
* Usually, any lion is majestic

The gardener planted roses in the garden.
I wrote Blair a letter and tore it up before I sent it.

**Train:**
The CoLA general acceptability corpus

**Test:**
NPI environment test sets

**Metric:**
Matthews Correlation (MCC) for acceptability
What if we train on NPI data directly?

*Those boys say that [the doctors ever went to an art gallery.]
*Those boys ever say that [the doctors went to an art gallery.]
Those boys say that [the doctors often went to an art gallery.]
Those boys often say that [the doctors went to an art gallery.]

*Who do you think that will question Seamus first?
*Usually, any lion is majestic.
The gardener planted roses in the garden.
I wrote Blair a letter, but I tore it up before I sent it.

Train:
The CoLA general acceptability corpus or NPI training set (hold-one-out by environment)

Test:
NPI environment test sets

Metric:
Matthews Correlation (MCC) for acceptability
What if we train on NPI data directly?

*Those boys say that [the doctors ever went to an art gallery.]*
*The gardener planted roses in the garden. I wrote Blair a letter. I tore it up before I sent it.*

**Train:**
The CoLA general acceptability corpus or NPI training set (hold-one-out by environment)

**Test:**
NPI environment test sets

**Metric:**
Matthews Correlation (MCC) for acceptability

BERT knows something about NPIs, but not all that much.
Let’s re-structure our data to isolate BERT’s knowledge of NPIs...

(1) Mary hasn’t eaten *any* cookies.
(2) *Mary has eaten *any* cookies.

**Train:**
The CoLA general acceptability corpus or NPI training set (hold-one-out by environment)

**Test:**
NPI environment test sets

**Metric:**
Pair accuracy over acceptability: How often does the model label both versions of a sentence correctly?
Let’s re-structure our data to isolate BERT’s knowledge of NPIs...

BERT knows something about NPIs, but not all that much.

(2) Mary has eaten any cookies.

**Train:**
The CoLA general acceptability corpus or NPI training set (hold-one-out by environment)

**Test:**
NPI environment test sets

**Metric:**
Pair accuracy over acceptability: How often does the model label both versions of a sentence correctly?
Let’s re-structure our data to isolate BERT’s knowledge of NPIs…

(1) Mary hasn’t eaten any cookies.

(2) *Mary has eaten any cookies.

**Train:**
The CoLA general acceptability corpus or NPI training set (hold-one-out by environment)

**Test:**
NPI environment test sets

**Metric:**
Pair preference accuracy: How often does the model assign a higher probability of acceptability to the correct sentence?
Let’s re-structure our data to isolate BERT’s knowledge of NPIs…

BERT has complete and perfect knowledge of NPI licensing.

(2) Mary has eaten any cookies.

**Train:**
The CoLA general acceptability corpus or NPI training set (hold-one-out by environment)

**Test:**
NPI environment test sets

**Metric:**
Pair preference accuracy: How often does the model assign a higher probability of acceptability to the correct sentence?
What if we ask BERT directly?

(1) Mary hasn’t eaten any cookies.
(2) *Mary has eaten any cookies.

Train:
The CoLA general acceptability corpus or NPI training set (hold-one-out by environment) or use BERT’s language modeling head directly

Test:
NPI environment test sets

Metric:
Pair preference accuracy: How often does the model assign a higher probability of acceptability to the correct sentence?

CoLA Training
NPI Training
MLM (No Training)
What if we ask BERT directly?

BERT does better than chance (50%), but not especially well.

(2) Mary has eaten *any* cookies.

**Train:**
- The CoLA general acceptability corpus
- or NPI training set (hold-one-out by environment)
- or use BERT’s language modeling head directly

**Test:**
- NPI environment test sets

**Metric:**
Pair preference accuracy: How often does the model assign a higher probability of acceptability to the correct sentence?
What if we use probing classifiers?

- Train: Scope prediction task, training only a small classifier without fine-tuning BERT (hold-one-out over environments)
- Test: Scope prediction task
- Metric: Matthews Correlation (MCC) for scope judgment
What if we use probing classifiers?

BERT knows a bit about NPIs, but it's not perfect.

**Train:**
Scope prediction task, training only a small classifier without fine-tuning BERT (hold-one-out over environments)

**Test:**
Scope prediction task

**Metric:**
Matthews Correlation (MCC) for scope judgment
<table>
<thead>
<tr>
<th></th>
<th>GloVe Bag-of-Words</th>
<th>BERT</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>75%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Those boys ever* say that [the doctors went to an art gallery.]

*Those boys say that* [the doctors often went to an art gallery.]

*Those boys only say that* [the doctors went to an art gallery.]

---

**Train:**
Scope prediction task, training BERT only (hold-out environments)

**Test:**
Scope prediction task

**Metric:**
Matthews Correlation (MCC) for scope judgment

---

What if we use probing classifiers?

---

BERT knows a bit about NPIs, but it's not perfect.

---

BERT knows something about NPIs, but not all that much.

---

BERT has complete and perfect knowledge of NPI licensing.

---

BERT does better than chance, but not especially well.
Back to evaluation...
There are plenty of big open problems in NLU, but doesn’t seem possible to build another GLUE-style benchmark again soon.

- Is our ability to build models improving faster than our ability to build hard evaluation sets?
Evaluation: What’s Next?

Give up and work on something else?

• I guess?

• or...
Evaluation: What’s Next?

Use adversarial filtering to semi-automatically create datasets that are hard for SotA models?

- Good source of data for training...
- Okay source of data for local hill-climbing evaluation...
Use adversarial filtering to semi-automatically create datasets that are hard for SotA models?

- Good source of data for training...
- Okay source of data for local hill-climbing evaluation...
- ...but using these datasets as benchmarks risks encouraging models that are different but not better.
- Mitigated by fast iteration times, but logistics get complicated.
Build *growing* benchmarks like Build-it-Break-it or ORB, where experts can add test data to target weaknesses.

- Similar risks, though to a lesser degree.
- Some risk that we lose sight of the task we're trying to solve.
Restrict the task training sets, or focus on zero-shot or few-shot adaptation to new tasks.

- Likely to encourage good representations…
- …but may not reflect the setting that we’re interested in.
Evaluation: What’s Next?

Build big, high-quality datasets?

- Aim for *hard* examples with human performance >99%.
- Doable! But slow, expensive, risky work.
One More Open Question

Is it possible to build benchmarks for bias that are robust and realistic enough that it’s worthwhile to hill-climb on them?
Evaluation: What’s Next?
Thanks!