Task-Independent Language Understanding
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.
Case Study: ELMo

Train large forward and backward deep LSTM language models.

Peters et al. '18
Case Study: ELMo

Train large (~100m-param) forward and backward deep LSTM language models.
Case Study: ELMo

Train large (~100m-param) forward and backward deep LSTM language models.
Case Study: ELMo

Train large (~100m-param) forward and backward deep LSTM language models.

This is a short sentence.
Case Study: ELMo

Peters et al. '18

Baseline

SQuAD | SNLI | SRL | Coref | NER | SST-5

40  | 85  | 70  | 55   | 100 | 40
Case Study: ELMo

Best paper at NAACL 2018!

Baseline
Baseline + ELMo

<table>
<thead>
<tr>
<th>Task</th>
<th>Baseline</th>
<th>Baseline + ELMo</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNLI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coref</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SST-5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Peters et al. '18
The Rest of the Talk

• The GLUE language understanding benchmark
  Wang et al. '18

  • ...and successes with unsupervised pretraining and
    fine-tuning on GLUE
  Radford et al. '18 (OpenAI GPT), Devlin et al. '18 (BERT)

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

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

• Easy transfer learning with STILTs
  Phang et al. '19, Wang et al. '19b
GLUE: What is it?
The General Language Understanding Evaluation (GLUE):

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

- Nine English-language sentence understanding tasks based on existing data, varying in:
  - Task difficulty
  - Training data volume and degree of training set–test set similarity
  - Language style/genre
  - Simple task APIs: All sentence or sentence-pair classification.
  - Simple leaderboard API: Upload predictions for a test set (Kaggle-style)
  - Usable with any kind of method/model!
## 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><strong>Single-Sentence Tasks</strong></td>
<td></td>
<td></td>
<td></td>
<td></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><strong>Similarity and Paraphrase Tasks</strong></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><strong>391k</strong></td>
<td>paraphrase</td>
<td>acc./F1</td>
<td>social QA questions</td>
</tr>
<tr>
<td><strong>Inference Tasks</strong></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>WNL1</td>
<td>634</td>
<td>71</td>
<td><strong>146</strong></td>
<td>coreference/NLI</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>

Wang, Singh, Michael, Hill, Levy & Bowman '18
## 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>634</td>
<td>71</td>
<td>146</td>
<td>coreference/NLI</td>
<td>acc.</td>
<td>fiction books</td>
</tr>
</tbody>
</table>
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 '18
Multi-Genre Natural Language Inference (MNLI)

Williams et al. '18

- Balanced classification for pairs of sentences into entailment, contradiction, and neutral
- Training set sentences drawn from five written and spoken genres. Dev/test sets divided into a matched set and a mismatched set with five more.

P: *The Old One always comforted Ca'daan, except today.*
H: *Ca'daan knew the Old One very well.*
neutral

<table>
<thead>
<tr>
<th>Corpus</th>
<th>MRPC</th>
<th>CoLA</th>
<th>SST-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.7k</td>
<td>408</td>
<td>1.7k</td>
</tr>
<tr>
<td></td>
<td>7k</td>
<td>1.5k</td>
<td>1.4k</td>
</tr>
<tr>
<td></td>
<td>364k</td>
<td>40k</td>
<td>391k</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inference Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNLI</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>QNLI</td>
</tr>
<tr>
<td>RTE</td>
</tr>
<tr>
<td>WNLI</td>
</tr>
</tbody>
</table>

Wang, Singh, Michael, Hill, Levy & Bowman '18
GLUE: What methods work?
Overall GLUE Score

GloVe BoW
Single-Task Model
Sentence-to-Vector
ELMo

Wang, Singh, Michael, Hill, Levy & Bowman '18
OpenAI’s GPT Language Model

• Same basic idea as ELMo, but many changes, including:
  
  • *Transformer* encoder architecture.
  
  • Entire network is *fine-tuned* for each task; few new parameters are added.

Radford et al. '18
OpenAI’s GPT Language Model

- Same basic idea as ELMo, but many changes, including:
  - *Transformer* encoder architecture.
  - Entire network is *fine-tuned* for each task; few new parameters are added.
  - Pretraining is on long spans of running text, not just isolated sentences.

Radford et al. '18
GLUE Score

- GloVe BoW
- Single-Task Model
- Sentence-to-Vector
- ELMo

Radford et al. '18
Google's BERT

Devlin et al. '18
see Baevski et al. '19 for similar concurrent work
The BERT Model

- Same basic idea as GPT with several changes, including:
  - Two different unlabeled data tasks in place of language modeling.
  - These allow the model to process both directions together with the same network at training time.
  - Bigger (100M => 300M params).

Devlin et al. '18

see Baevski et al. '19 for similar concurrent work
GLUE Score

- GloVe BoW
- Single-Task Model
- Sentence-to-Vector
- ELMo
- OpenAI GPT

Devlin et al. '18

see Baevski et al. '19 for similar concurrent work
see Baevski et al. '19 for similar concurrent work
see Baevski et al. '19 for similar concurrent work
Why does BERT work so well?
What does BERT know?
Edge Probing

Pre-trained encoder (ELMo, BERT, etc.)

MLP

0 1 0 0 ... 

<A0> <A1> <A2> <A3> ...

Labels

Binary classifiers

Span representations

Contextual vectors

Input tokens

Tenney, Xia, Chen, Wang, Poliak, McCoy, Kim, Van Durme, Bowman, Das, & Pavlick '19
Edge Probing

Pre-trained encoder (ELMo, BERT, etc.)

Parts of speech? Sentence structure? Entity names? Coreference?

MLP

[1,2) [2,5)

Pre-trained encoder (ELMo, BERT, etc.)

Input tokens

Labels
Binary classifiers
Span representations
Contextual vectors

Tenney, Xia, Chen, Wang, Poliak, McCoy, Kim, Van Durme, Bowman, Das, & Pavlick '19
Edge Probing with ELMo

- Parts of speech?
- Sentence structure?
- Entity names?
- Coreference?
Edge Probing with ELMo and BERT

- Parts of speech?
- Sentence structure?
- Entity names?
- Coreference?
How much can we trust our conclusions?
How much can we trust these conclusions?

- Studies like ours that use auxiliary analysis datasets 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 a few 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,
  
  - Common in natural data.

  - Well-characterized in the linguistics literature.

  - Depends on long-distance dependencies and complex structures, rather than local co-occurrence.

  - Should be learnable from raw text alone.

- *Does BERT know when NPIs are licensed?*

  (1) Mary hasn’t eaten any cookies.

  (2) *Mary has eaten any cookies.*
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,

- Common in natural data.

- Well-characterized in the linguistics literature.

- Depends on long-distance dependencies and complex structures, rather than local co-occurrence.

- Should be learnable from raw text alone.

- *Does BERT know when NPIs are licensed?*

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

(1) Mary hasn’t eaten *any* cookies.

(2) *Mary has eaten *any* cookies.

{Warstadt, Cao, Grosu, Peng, Blix, Nie, Alsop, Bordia, Liu, Parrish, Wang, Phang, Mohananey, Htut, Jeretič} & Bowman ‘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 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
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:
NPI training set (hold-one-out by environment)
or 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.]

CoLA Training

NPI Training

GloVe Bag-of-Words

BERT

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

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...

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

Test:
- NPI environment test sets

Metric:
Pair accuracy over acceptability: How often does the model label both versions of a sentence correctly?

(1) Mary hasn’t eaten any cookies.
(2) *Mary has eaten any cookies.
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:**
NPI training set (hold-one-out by environment)
or the CoLA general acceptability corpus

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

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

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

![Bar chart](image)
- GloVe Bag-of-Words
- BERT
What if we ask BERT directly?

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

(2) "Mary has eaten any cookies.

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

Those boys wonder *whether* [the doctors *ever* went to an art gallery.]

Those boys *ever* wonder *whether* [the doctors went to an art gallery.]

Those boys wonder *whether* [the doctors *often* went to an art gallery.]

Those boys *often* wonder *whether* [the doctors went to an art gallery.]

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.]

---

**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
What if we use probing classifiers?

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

BERT does better than chance, but not especially well.

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

BERT has complete and perfect knowledge of NPI licensing.

BERT knows something about NPIs, but not all that much.
Recent Progress on GLUE
Building a Better Muppet

- Lots of follow-up work, including:
  - MT-DNN/ALICE: Multi-task fine-tuning; ensembling
  - RoBERTa: Simplified objective; more training data
  - ALBERT: Modified objective; parameter sharing across layers

Liu et al. '19a, Wang et al. '19, Liu et al. '19b, Anonymous '19
GLUE Score

- GloVe BoW
- Single-Task Model
- Sentence-to-Vector
- ELMo
- OpenAI GPT
- BERT Large
GLUE Score

Liu et al. '19
GLUE Score

- GloVe BoW
- Single-Task Model
- Sentence-to-Vector
- ELMo
- OpenAI GPT
- BERT Large
- MT-DNN
- RoBERTa

Liu et al. '19
• How much headroom does GLUE have left?

• To compute a conservative estimate for each task:
  
  • Train crowdworkers with instructions, plus twenty labeled development set examples in an interactive training mode.
  
  • Collect five labels per example for 500 test set examples.
GLUE Score

Nangia & Bowman '19
GLUE Score

GloVe BoW  Single-Task Model  Sentence-to-Vector  ELMo  OpenAI GPT  BERT Large  Human Baseline  MT-DNN  RoBERTa  ALBERT

Nangia & Bowman '19
A revised version of GLUE with:

- A new set of eight target tasks...
- ...selected from 30+ submissions to an open call for participation to be easy for humans and hard for BERT.
- A slightly expanded set of task APIs (including multiple-choice QA, word-in-context classification, and more)

{Wang, Pruksachatkun, Nangia, Singh}, Michael, Hill, Levy & Bowman '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>
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 **Entailment:** Unknown

<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</td>
<td>{Wang, Pruksachatkun, Nangia, Singh}, Michael, Hill, Levy &amp; Bowman '19</td>
</tr>
<tr>
<td>ReCoRD</td>
<td>101k</td>
<td>10k</td>
<td>10k</td>
<td>QA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MultiRC
Khashabi et al. '18

- Multiple choice reading comprehension QA over paragraphs.

**Paragraph:** (CNN) – Gabriel García Márquez, widely regarded as one of the most important contemporary Latin American authors, was admitted to a hospital in Mexico earlier this week, according to the Ministry of Health. The Nobel Prize recipient, known as “Gabo,” had infections in his lungs and his urinary tract. He was suffering from dehydration, the ministry said. García Márquez, 87, is responding well to antibiotics, but his release date is still to be determined. “I wish him a speedy recovery.” Mexican President Enrique Peña wrote on Twitter. García Márquez was born in the northern Colombian town of Aracataca, the inspiration for the fictional town of Macondo, the setting of the 1967 novel “One Hundred Years of Solitude.” He won the Nobel Prize for literature in 1982 “for his novels and short stories, in which the fantastic and the realistic are combined in a richly composed world of imagination, reflecting a continent’s life and conflicts,” according to the Nobel Prize website. García Márquez has spent many years in Mexico and has a huge following there. Colombian President Juan Manuel Santos said his country is thinking of the author. “All of Colombia wishes a speedy recovery to the greatest of all time: Gabriel García Márquez,” he tweeted. CNN en Español’s Fidel Gutiérrez contributed to this story.

**Question:** Whose speedy recover did Mexican President Enrique Peña wish on Twitter?

**Candidate answers:** Enrique Peña (F), Gabriel Garcia Marquez (T), Gabo (T), Gabriel Mata (F), Fidel Gutiérrez (F), 87 (F), The Nobel Prize recipient (T)

<table>
<thead>
<tr>
<th>COPA</th>
<th>MultiRC</th>
<th>ReCoRD</th>
<th>RTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc.</td>
<td>5100</td>
<td>101k</td>
<td>2500</td>
</tr>
<tr>
<td>blogs, photography encyclopedia</td>
<td>953</td>
<td>10k</td>
<td>278</td>
</tr>
<tr>
<td>QA</td>
<td>1800</td>
<td>10k</td>
<td>300</td>
</tr>
<tr>
<td>F1a/EM</td>
<td>various</td>
<td>I</td>
<td>W {Wang, Pruksachatkun, Nangia, Singh}, Michael, Hill, Levy &amp; Bowman '19</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>$F_{1a}/EM$</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 '19
SuperGLUE Score

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe Bag of Words</td>
<td>45</td>
</tr>
<tr>
<td>BERT</td>
<td>60</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>82.5</td>
</tr>
<tr>
<td>Human Estimate</td>
<td>95</td>
</tr>
</tbody>
</table>

{Wang, Pruksachatkun, Nangia, Singh}, Michael, Hill, Levy & Bowman '19
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 use some naturally occurring and 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), RoBERTa ~9x more sensitive to irrelevant gender information than humans.
A Handy Trick
• What if you want to solve a hard task with limited training data, but have access to abundant data for another task with that uses similar skills?

• Example: Commitment Bank (250) with MNLI (393k)

• Supplementary Training on Intermediate Labeled-data Tasks (STILTs) is an easy but very robust solution:
  • Download a large model like BERT that was pretrained on unlabeled data.
  • Fine tune that model on the intermediate labeled-data task.
  • Fine tune the same model further on the target task.

Phang, Févry & Bowman '18
BERT on STILTs

• +1.5 on GLUE w/ MNLI and QQP
• +2.5 on SuperGLUE w/ MNLI

Clark et al. '19: +3.7 on BoolQ w/ MNLI

Sap et al. '18: +4 to +8 on commonsense tasks w/ SocialIQA

MNLI+STILTs built into RoBERTa and ALBERT
BERT on STILTs

- +1.5 on GLUE w/ MNLI and QQP
- +2.5 on Sap et al. '18 w/ MNLI
- Clark et al. '19: +3.7 on BoolQ w/ MNLI
- Sap et al. '18: +4 to +8 on commonsense tasks w/ SocialIQA
- MNLI+STILTs built into RoBERTa and ALBERT

Tuning Not Required!
## ELMo and BERT Base on STILTs

<table>
<thead>
<tr>
<th>Intermediate Task</th>
<th>Avg</th>
<th>CoLA</th>
<th>SST</th>
<th>ELMo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random$_{g}^{+}$</td>
<td>70.5</td>
<td>38.5</td>
<td>87.7</td>
<td></td>
</tr>
<tr>
<td>Single-Task$_{g}^{+}$</td>
<td>71.2</td>
<td>39.4</td>
<td>90.6</td>
<td></td>
</tr>
<tr>
<td>CoLA$_{g}^{+}$</td>
<td>71.1</td>
<td>39.4</td>
<td>87.3</td>
<td></td>
</tr>
<tr>
<td>SST$_{g}^{+}$</td>
<td>71.2</td>
<td>38.8</td>
<td>90.6</td>
<td></td>
</tr>
<tr>
<td>MRPC$_{g}^{+}$</td>
<td>71.3</td>
<td>40.0</td>
<td>88.4</td>
<td></td>
</tr>
<tr>
<td>QQP$_{g}^{+}$</td>
<td>70.8</td>
<td>34.3</td>
<td>88.6</td>
<td></td>
</tr>
<tr>
<td>STS$_{g}^{+}$</td>
<td>71.6</td>
<td>39.8</td>
<td>88.4</td>
<td></td>
</tr>
<tr>
<td>MNLI$_{g}^{+}$</td>
<td>72.1</td>
<td>38.9</td>
<td>89.0</td>
<td></td>
</tr>
<tr>
<td>QNLI$_{g}^{+}$</td>
<td>71.2</td>
<td>37.2</td>
<td>88.3</td>
<td>81.1/69</td>
</tr>
<tr>
<td>RTE$_{g}^{+}$</td>
<td>71.2</td>
<td>38.5</td>
<td>87.7</td>
<td>81.1/87</td>
</tr>
<tr>
<td>WNLI$_{g}^{+}$</td>
<td>71.9</td>
<td>38.4</td>
<td>88.6</td>
<td>78.4/55.9</td>
</tr>
<tr>
<td>DisSent WPI$_{g}^{+}$</td>
<td>72.1</td>
<td>39.9</td>
<td>87.6</td>
<td>81.9/87</td>
</tr>
<tr>
<td>MT En-Eu$_{g}^{+}$</td>
<td>72.1</td>
<td>40.1</td>
<td>87.8</td>
<td>79.9/86.6</td>
</tr>
<tr>
<td>MT En-Ru$_{g}^{+}$</td>
<td>70.4</td>
<td>41.0</td>
<td>86.9</td>
<td>76.6/86.5</td>
</tr>
<tr>
<td>Reddit$_{g}^{+}$</td>
<td>71.0</td>
<td>38.5</td>
<td>87.7</td>
<td>77.2/85.0</td>
</tr>
<tr>
<td>SkipThought$_{g}^{+}$</td>
<td>71.7</td>
<td>40.6</td>
<td>87.7</td>
<td>79.7/86.5</td>
</tr>
<tr>
<td>MTL GLUE$_{g}^{+}$</td>
<td>72.1</td>
<td>33.8</td>
<td>90.5</td>
<td>81.1/87.4</td>
</tr>
<tr>
<td>MTL Non-GLUE$_{g}^{+}$</td>
<td>72.4</td>
<td>39.4</td>
<td>88.8</td>
<td>80.6/86.8</td>
</tr>
<tr>
<td>MTL All$_{g}^{+}$</td>
<td>72.2</td>
<td>37.9</td>
<td>89.6</td>
<td>79.2/86.4</td>
</tr>
</tbody>
</table>

- Most intermediate tasks harm performance, especially with BERT.
- This includes most of the GLUE tasks, MT, Reddit prediction, DisSent, and several more!
- BERT with MNLI or BERT with GLUE (multi-task) work best, and show consistent improvements.

---

Wang, Hula, Xia, Pappagari, McCoy, Patel, Kim, Tenney, Huang, Yu, Jin, Chen, Van Durme, Grave, Pavlick and Bowman '19
Practical Conclusions

• If you’re building a language understanding model now, you have at least a few thousand training examples, and you need the best performance you can get:

• Use RoBERTa.

• If you're aware of a big dataset for some related task, or if you're working with very limited training data, use STILTs, too!

• Don’t be too quick to trust any one analysis study that claims to tell you what NLP models know.

• Keep an eye on super.gluebenchmark.com for future developments in this area.

• For a toolkit that implements everything I've spoken about, try jiant.info.
Open Questions

Plenty of open questions!

• How far can we push plain unsupervised pretraining with bigger models?

• What makes a task suitable for use as an intermediate task in STILTs?

• Are we nearing the end of the line for evaluation with IID test sets?

• How can we mitigate the social biases that these models learn during pretraining and fine-tuning?
Thanks!

Questions:
bowman@nyu.edu

@sleepinyourhat

Try SuperGLUE:
super.gluebenchmark.com
Sponsors

See cited papers for full project details.
But wait! There's more!
<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Model</th>
<th>URL</th>
<th>Score</th>
<th>BoolIQ</th>
<th>CB</th>
<th>COPA</th>
<th>MultiRC</th>
<th>ReCoRD</th>
<th>RTE</th>
<th>WiC</th>
<th>WSC</th>
<th>AX-b</th>
<th>AX-g</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SuperGLUE Human Baselines</td>
<td>SuperGLUE Human Baselines</td>
<td>89.8</td>
<td>89.0</td>
<td>95.8/98.9</td>
<td>100.0</td>
<td>81.8/51.9</td>
<td>91.7/91.3</td>
<td>93.6</td>
<td>80.0</td>
<td>100.0</td>
<td>76.6</td>
<td>99.3/99.7</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Facebook AI</td>
<td>RoBERTa</td>
<td>84.6</td>
<td>87.1</td>
<td>90.5/95.2</td>
<td>90.6</td>
<td>84.4/52.5</td>
<td>90.6/90.0</td>
<td>88.2</td>
<td>69.9</td>
<td>89.0</td>
<td>57.9</td>
<td>91.0/78.1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>SuperGLUE Baselines</td>
<td>BERT++</td>
<td>71.5</td>
<td>79.0</td>
<td>84.8/90.4</td>
<td>73.8</td>
<td>70.0/24.1</td>
<td>72.0/71.3</td>
<td>79.0</td>
<td>69.6</td>
<td>64.4</td>
<td>38.0</td>
<td>99.4/51.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td></td>
<td>69.0</td>
<td>77.4</td>
<td>75.7/83.6</td>
<td>70.6</td>
<td>70.0/24.1</td>
<td>72.0/71.3</td>
<td>71.7</td>
<td>69.6</td>
<td>64.4</td>
<td>23.0</td>
<td>97.8/51.7</td>
<td></td>
</tr>
</tbody>
</table>