Natural Language Representations and Challenges

Slav Petrov
on behalf of the
Language Team @ Google Research
Outline

● Representations
  ○ BERT
  ○ Relations
  ○ Paragraphs

● Challenges
  ○ Natural Questions
  ○ Yes/No Questions
  ○ Identifying Commands
State-of-the-art in Natural Language Understanding in 2017

Features → Neural Networks → Custom (Recurrent) Architectures

Question Answering

Inference

Named Entity Recognition

Seo, Kembhavi, Farhadi, Hajishirzi, ICLR 2017

Chen, Zhu, Ling, Wei, Jiang, Inkpen, ACL 2017

Peters, Neumann, Iyyer, Gardner, Clark, Lee, Zettlemoyer, ACL 2017
Oct. 2018: One Model with Task-specific Tuning in Minutes

(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

NAACL 2019

Best Long Paper

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova
# Question Answering (SQuAD 1.1)

As of yesterday (July 12)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BERT + DAE + AoA (ensemble)</td>
<td>87.147</td>
<td>89.474</td>
</tr>
<tr>
<td>2</td>
<td>BERT + ConvLSTM + MTL + Verifier (ensemble)</td>
<td>86.730</td>
<td>89.286</td>
</tr>
<tr>
<td>3</td>
<td>BERT + N-Gram Masking + Synthetic Self-Training (ensemble)</td>
<td>86.673</td>
<td>89.147</td>
</tr>
<tr>
<td>4</td>
<td>XLNet (single model)</td>
<td>86.346</td>
<td>89.133</td>
</tr>
<tr>
<td>5</td>
<td>SemBERT (ensemble)</td>
<td>86.166</td>
<td>88.886</td>
</tr>
<tr>
<td>6</td>
<td>BERT + DAE + AoA (single model)</td>
<td>85.884</td>
<td>88.621</td>
</tr>
</tbody>
</table>

**Top Performing Models**

1. **BERT + DAE + AoA (ensemble)**
   - Stanford University
   - EM: 86.831
   - F1: 89.452

2. **BERT + ConvLSTM + MTL + Verifier (ensemble)**
   - Layer 6 AI
   - EM: 86.730
   - F1: 89.286

3. **BERT + N-Gram Masking + Synthetic Self-Training (ensemble)**
   - Google AI Language
   - EM: 86.673
   - F1: 89.147

**April 5**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>BERT + DAE + AoA (single model)</td>
<td>85.884</td>
<td>88.621</td>
</tr>
</tbody>
</table>

**Top Performing Models**

4. **BERT + DAE + AoA (single model)**
   - Joint Laboratory of HIT and iFLYTEK Research
   - EM: 85.884
   - F1: 88.621

5. **BERT + MML + ADA (ensemble)**
   - Microsoft Research Asia
   - EM: 85.082
   - F1: 87.615

6. **BERT + ConvLSTM + MTL + Verifier (single model)**
   - Layer 6 AI
   - EM: 84.924
   - F1: 88.204
Pre-Training in NLP

- Use co-occurrence statistics to learn word embeddings:

  - Inner Product
  - Inner Product

  king wore a crown ← king
  queen wore a crown ← queen

  \([-0.5, -0.9, 1.4, \ldots]\)

  \([-0.6, -0.8, -0.2, \ldots]\)

- Problem: Word embeddings are applied in a context free manner:

  - open a bank account
  - on the river bank

  \([0.3, 0.2, -0.8, \ldots]\)
History of Contextual Representations

- Semi-Supervised Sequence Learning, Google, 2015
History of Contextual Representations

- ELMo: Deep Contextual Word Embeddings, AI2 & University of Washington, 2017

Train Separate Left-to-Right and Right-to-Left LMs

Apply as “Pre-trained Embeddings”
History of Contextual Representations

- Improving Language Understanding by Generative Pre-Training, OpenAI, 2018
Unidirectional vs. Bidirectional Models

**Unidirectional context**
Build representation incrementally

**Bidirectional context**
Words can “see themselves”
Masked Language Model (Fill-in-the-blank)

- **Solution**: Mask out $k\%$ of the input words, and then predict the masked words
  - We always use $k = 15\%$

  the man went to the [MASK] to buy a [MASK] of milk

- Too little masking: Too expensive to train
- Too much masking: Not enough context
Next Sentence Prediction

- To learn relationships between sentences, predict whether Sentence B is an actual sentence that proceeds Sentence A, or a random sentence.

<table>
<thead>
<tr>
<th>Sentence A</th>
<th>Sentence B</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>The man went to the store.</td>
<td>He bought a gallon of milk.</td>
<td>IsNextSentence</td>
</tr>
<tr>
<td>The man went to the store.</td>
<td>Penguins are flightless.</td>
<td>NotNextSentence</td>
</tr>
</tbody>
</table>
Next Sentence Prediction

- Use 30,000 WordPiece vocabulary on input
- Each token is sum of three embeddings
- Single sequence is much more efficient
Transformer Architecture

- **Multi-headed self attention**
  - Models context

- **Feed-forward layers**
  - Computes non-linear hierarchical features

- **Layer norm and residuals**
  - Makes training deep networks healthy

- **Positional embeddings**
  - Allows model to learn relative positioning
From One-Hot Vectors to Word Embeddings & Self-Attention

The Annotated Transformer, The Illustrated Transformer, The Illustrated BERT
From One-Hot Vectors to Word Embeddings & Self-Attention

The Annotated Transformer, The Illustrated Transformer, The Illustrated BERT
From One-Hot Vectors to Word Embeddings & Self-Attention

The Annotated Transformer, The Illustrated Transformer, The Illustrated BERT
From One-Hot Vectors to Word Embeddings & Self-Attention

The Annotated Transformer, The Illustrated Transformer, The Illustrated BERT
From One-Hot Vectors to Word Embeddings & Self-Attention

one-hot    embedding    query, key, value    (self-)attention

contextual representation for the word bank

The Annotated Transformer, The Illustrated Transformer, The Illustrated BERT
Transformer vs LSTM

- Self-attention == no locality bias
  - Long-distance context has “equal opportunity”

- Single multiplication per layer == efficiency on TPU
  - Effective batch size is number of words, not sequences
Basic BERT Recipe

Pre-Training

BERT

Masked Sentence A  Masked Sentence B
Basic BERT Recipe

Pre-Training

Fine-Tuning

SQuAD

Start/End Span

[Diagram of BERT pre-training and fine-tuning processes with NSP, Mask LM, and Mask LM inputs.]
### MultiNLI
- **Premise**: Hills and mountains are especially sanctified in Jainism.
- **Hypothesis**: Jainism hates nature.
- **Label**: Contradiction

### CoLa
- **Sentence**: The wagon rumbled down the road.
- **Label**: Acceptable
- **Sentence**: The car honked down the road.
- **Label**: Unacceptable
Results: Commonsense Reasoning and Question Answering

On stage, a woman takes a seat at the piano. She
a) sits on a bench as her sister plays with the doll.
b) smiles with someone as the music plays.
c) is in the crowd, watching the dancers.
d) nervously sets her fingers on the keys.

How far is Jacksonville from Miami?

Ground Truth Answers: 340 miles 340 miles 340 miles
Prediction: 340 miles

Jacksonville is in the First Coast region of northeast Florida and is centered on the banks of the St. Johns River, about 25 miles (40 km) south of the Georgia state line and about 340 miles (550 km) north of Miami. The Jacksonville Beaches communities are along the adjacent Atlantic coast. The area was originally inhabited by the Timucua people, and in 1564 was the site of the French colony of Fort Caroline, one of the earliest European settlements in what is now the continental United States. Under British rule, settlement grew at the narrow point in the river where cattle crossed, known as Wacca Pilatka to the Seminole and the Cow Ford to the British. A platted town was established there in 1822, a year after the United States gained Florida from Spain; it was named after Andrew Jackson, the first military governor of the Florida Territory and seventh President of the United States.

<table>
<thead>
<tr>
<th>System</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESIM+GloVe</td>
<td>51.9</td>
<td>52.7</td>
</tr>
<tr>
<td>ESIM+ELMo</td>
<td>59.1</td>
<td>59.2</td>
</tr>
<tr>
<td>BERT\textsubscript{BASE}</td>
<td>81.6</td>
<td>-</td>
</tr>
<tr>
<td>BERT\textsubscript{LARGE}</td>
<td>86.6</td>
<td>86.3</td>
</tr>
<tr>
<td>Human (expert)(^\d)</td>
<td>-</td>
<td>85.0</td>
</tr>
<tr>
<td>Human (5 annotations)(^\d)</td>
<td>-</td>
<td>88.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BERT (ensemble)</td>
<td>87.433</td>
<td>93.160</td>
</tr>
<tr>
<td></td>
<td>Stanford University (Rajpurkar et al. '16)</td>
<td>82.304</td>
<td>91.221</td>
</tr>
<tr>
<td></td>
<td>Human Performance</td>
<td>82.304</td>
<td>91.221</td>
</tr>
</tbody>
</table>

https://arxiv.org/abs/1810.04805
Ablation Experiments

- Masked LM (compared to left-to-right LM) is very important on some tasks, Next Sentence Prediction is important on other tasks.
- Left-to-right model does very poorly on word-level task (SQuAD), although this is mitigated by BiLSTM
More is Better

- More data (and training longer) helps => not yet asymptoted
- Bigger model helps a lot
Try It Out, Get Faster Training with TPUs

BERT finetuning tasks in 5 minutes with Cloud TPU

BERT, or Bidirectional Embedding Representations from Transformers, is a new method of pre-training language representations which obtains state-of-the-art results on a wide array of Natural Language Processing (NLP) tasks. The academic paper can be found here: https://arxiv.org/abs/1810.04805.

This Colab demonstrates using a free Colab Cloud TPU to fine-tune sentence and sentence-pair classification tasks built on top of pretrained BERT models.

Note: You will need a GCP (Google Compute Engine) account and a GCS (Google Cloud Storage) bucket for this Colab to run.

Please follow the Google Cloud TPU quickstart for how to create GCP account and GCS bucket. You have $300 free credit to get started with any GCP product. You can learn more about Cloud TPU at https://cloud.google.com/tpu/docs.
Do I Need Full BERT Models for All My Tasks?

Tune only ~2% of the parameters per task

Large BERT layer *frozen*  Small Adapter layer *tuned*

Houlsby, Giurgiu, Jastrzebski, Morrone, de Laroussilhe, Gesmundo, Attariyan, Gelly, arxiv Feb 2019
Recently: Even Better Pretraining

SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human Performance</td>
<td>82.304</td>
<td>91.221</td>
</tr>
<tr>
<td></td>
<td>Stanford University</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Rajpurkar et al. '16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>XLNet (single model)</td>
<td>89.898</td>
<td>95.080</td>
</tr>
<tr>
<td></td>
<td>XLNet Team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BERT (ensemble)</td>
<td>87.433</td>
<td>93.160</td>
</tr>
<tr>
<td></td>
<td>Google AI Language</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

May 21, 2019

Oct 05, 2018
<table>
<thead>
<tr>
<th></th>
<th>BERT</th>
<th>XLNet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objective</strong></td>
<td>Masked LM + NextSentence</td>
<td>Autoregressive LM</td>
</tr>
<tr>
<td><strong>Masking</strong></td>
<td>Random 15%</td>
<td>Last ⅔ in permutation order</td>
</tr>
<tr>
<td><strong>Position Encoding</strong></td>
<td>Absolute</td>
<td>Relative</td>
</tr>
<tr>
<td><strong>Data</strong></td>
<td>13 GB of text</td>
<td>126 GB of text</td>
</tr>
</tbody>
</table>
Other BERT-inspired work

- What does BERT learn - Tenney et al., ACL 2019
- Relation learning - Baldini-Soares et al., ACL 2019
- Passage representations - Lee et al., ACL 2019
What does BERT know about language?

- What linguistic relationships?
- Where in the model are they computed?

**Classifier-based probing:** project model activations into space of linguistic annotations (graph edges)
BERT RedisCOVERs the Classical NLP Pipeline

High weights for POS in lower layers, then constituents, dependencies, and SRL, followed by entities and coreference as we move up the stack!

BERT improves coref over ELMo (84->91, or 39% relative)

We can trace hypotheses on individual sentences!

“he smoked **toronto** in the playoffs with six hits, seven walks ...”
Representing Entities and Relations

[BERT], inspired by Cale's earlier cover, recorded one of the most acclaimed versions of "[BLANK]"

[BERT]'s rendition of "[BLANK]" has been called "one of the great songs" by Time, ...

In 1997, [BLANK] moved to [BLANK], to resume work on the album

Baldini-Soares et al., ACL 2019
Matching The Blanks Results

FewRel is a **few-shot relation extraction** task, for which new relations are seen at test time.

Given a single exemplar, choose which of 10 other sentences express the same relation.

Matching the Blanks (MTB) beats best published system *without any of FewRel’s training data*.

Both standard BERT and BERT trained with MTB easily beat previous work, but with MTB we need many fewer labels.

Baldini-Soares et al., ACL 2019
Open Retrieval Question Answering

**Input**
- What does the zip in zip code stand for?
- How many districts are in Alabama?

**Latent Retrieval**
- Wikipedia

**Output**
- Zone Improvement Plan
- 7

**Goal:** Learn to efficiently read Wikipedia *without any retrieval data.*

**Motivation:** Best known recipe for latent retrieval is TF-IDF filtering + brute force. We can do better.

**Key Insight:** Pre-train an unsupervised ScaM neural retriever. This enables efficient end-to-end fine-tuning with standard latent variable learning methods.
Zebras have four gaits: walk, trot, canter and gallop. They are generally slower than horses, but their great stamina helps them outrun predators. When chased, a zebra will zig-zag from side to side, making it more difficult for the predator to attack...

In-batch negative example
Poe capitalized on the success of "The Raven" by following it up with his essay "The Philosophy of Composition" (1846), in which he detailed the poem's creation...

Pseudo Evidence
Zebras have four gaits: walk, trot, canter and gallop. When chased, a zebra will zig-zag from side to side, making it more difficult for the predator to attack...

In-batch negative example
Gagarin was further selected for an elite training group known as the Sochi Six, from which the first cosmonauts of the Vostok programme would be chosen...

Progress towards one of the hardest binary text classification tasks today:

Can you be charged for the same crime in two different states?

Wikipedia

TF-IDF + BERT
ICT + BERT

Natural Questions
WebQuestions
CuratedTrec
Realistic Challenge Sets

Natural Questions - Kwiatkowski et al., TACL 2019
Yes/No Questions - Clark et al., NAACL 2019
Identifying Commands - Elkahky et al., EMNLP 2018
Natural Questions Motivation

**Goal:** Provide academia with first question answering dataset that represents a real question answering problem.

Previous question answering datasets are contrived. E.g. SQuAD’s questions often paraphrase evidence text.

Answering real user queries requires much deeper language understanding and world knowledge.

Many questions have multiple acceptable answers: *last hurricane in Massachusetts* has a formal meaning (eye of the storm in MA) and a different colloquial meaning (hurricane force winds in MA).

NQ embraces this acceptable variability. Solutions should model the full distribution of possible answers.

---

**Question:** The *success of* the Britain Can Make It exhibition *led to the planning of* what exhibition in 1951?

**Evidence:** ... The *success of* this exhibition *led to the planning of* the Festival of Britain (1951). By 1948 most of the collections had been returned to the museum.

**Answer:** Festival of Britain

---

**Question:** Can you make and receive calls in *airplane mode*?

**Evidence:** *Airplane mode*, .... suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.

**Answer:** No
Challenge: Many Correct Answers

NQ annotators are encouraged to pick the first good answer.

In practice we sometimes get many different answer locations for the same question.

**Question:** name the substance used to make the filament of bulb

Consumption of incandescent light bulbs grew rapidly in the US. When tungsten filaments were introduced, about 50 million lamp filaments, and by 1945, annual sales of lamps were 795 million (m). Lamp, the uncoiled length of the tungsten filament is usually 22.8 inches (580 mm), and the filament diameter is 0.0018 inches (0.046 mm). The incandescent light bulbs consist of an air-tight glass enclosure (the envelope, or bulb) with a filament of tungsten wire inside the bulb, through which an electric current is passed.
Defining Correctness

Wrong annotations are often the result of annotators trying to find an answer when the evidence isn’t sufficient.

When all annotators agree that there is enough evidence available to answer a question, the annotations are overwhelmingly correct.

**X-axis** - proportion of annotations that are non-null for question.  
**Y-axis** - expectation that a non-null annotation's question is in this bucket.

Also broken down, conditioned on annotation being: Correct \((C)\); Correct but debatable \((C_d)\); or Wrong \((W)\).
Natural Questions - Status

- First ever release of Google queries.
- 300k training items, 16k for evaluation. Upper bound of 87% on long answers, 76% on short.
- Leaderboard seeing good activity, task is quite a bit harder than Squad.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>Participant</th>
<th>Affiliation</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BERT-dm_v2-ensemble</td>
<td>DREAM</td>
<td>Anonymous</td>
<td>0.74498</td>
</tr>
<tr>
<td>2</td>
<td>bert-cs</td>
<td>wangxuguang</td>
<td>orionstar-nlp</td>
<td>0.72618</td>
</tr>
<tr>
<td>3</td>
<td>bert-dm</td>
<td>DREAM</td>
<td>Anonymous</td>
<td>0.72613</td>
</tr>
<tr>
<td>4</td>
<td>bert_dm</td>
<td>DREAM</td>
<td>Anonymous</td>
<td>0.7196</td>
</tr>
<tr>
<td>5</td>
<td>bert-dm</td>
<td>DREAM</td>
<td>Anonymous</td>
<td>0.70248</td>
</tr>
<tr>
<td>6</td>
<td>BERT</td>
<td>LEE_SJ</td>
<td>LAIR</td>
<td>0.68803</td>
</tr>
<tr>
<td>7</td>
<td>BERT-mnlp-ensemble</td>
<td>GAAMA</td>
<td>IBM Research AI</td>
<td>0.68702</td>
</tr>
</tbody>
</table>
Prior Approaches to Testing Inference/Reasoning Abilities

Write one sentence that is definitely correct about the situation or event in the line.

At 8:34, the Boston Center controller received a third transmission from American 11.

The Boston Center controller got a third transmission from American 11.
BoolQ: Naturally Occurring Yes-No Questions

Is You and I by Lady Gaga a cover?

Is there such thing as a dominant eye?

Is the Arch in St. Louis a national park?
Real Problems that Naturally Require Inference to Solve

**Question**: Did Bonnie Blair’s daughter make the Olympic team?

**Passage**: Blair and Cruikshank have two children: a son, Grant, and daughter, Blair....Blair Cruikshank competed at the 2018 United States Olympic speed skating trials at the 500 meter distance.

**Answer**: No
Collecting Passages

Are there blue whales in the Atlantic Ocean?

Yes

Occasionally blue whales get lost in the Baltic Sea. The remains of two blue whales have been identified in Finland. One 19-meter skeleton was found on the bottom of the Gulf of Finland during the construction of the Nord Stream pipeline. The vertebrae of a second individual were recovered from a field near Porvoo on the coast of the Bothnian Sea in 1942 and identified in the Copenhagen Zoological Museum by Magnus Degerbol. Radiocarbon dating revealed the former had wandered into the Baltic Sea over 7000 years ago and perished, while the latter bones were shown to be 800–900 years old.

Pipeline from Natural Questions (Kwiatkowski et al., 2019)
Test Set Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>62.31</td>
</tr>
<tr>
<td>Recurrent</td>
<td>67.52</td>
</tr>
<tr>
<td>Recurrent + MNLI</td>
<td>74.24</td>
</tr>
<tr>
<td>BERT</td>
<td>76.7</td>
</tr>
<tr>
<td>BERT + MNLI</td>
<td>80.43</td>
</tr>
</tbody>
</table>
Sample Efficiency: MultiNLI > BERT for Small Data
Noun-Verb Ambiguity

“lives” / Noun → Pronounced /lalvz/
“lives” / Verb → Pronounced /lIvz/

All earlier taggers get these wrong:

1. Certain insects can damage plumerias, such as mites, flies, or aphids. NOUN
2. Mark which area you want to distress. VERB
Certain insects can damage plumerias, such as mites, flies, or aphids. Mark which area you want to distress.

ML can be fooled by corpus frequency.

"A Challenge Set and Methods for Noun-Verb Ambiguity", EMNLP 2018

ACL 2018; won CoNLL 2018 shared task on universal part-of-speech tagging
Accuracy on Noun-Verb Disambiguation

Contextual Representations (Best Paper Award at NAACL 2018): Had showed no improvement on standard tagging benchmark.

- Bohnet et al. (2018): 74
- w/ELMo: 82.1
- w/NV Data: 86.4
- w/ELMo+NV Data: 88.9
Released Datasets with “In-the-Wild” Natural Challenges

Are there blue whales in the Atlantic Ocean? YES

Occasionally blue whales get lost in the Baltic Sea. The remains of two blue whales have been identified in Finland. One 10-meter skeleton was found on the bottom of the Gulf of Finland during the construction of Nord Stream pipeline. The vertebrae of a second individual were recovered from a field near the coast of Sweden. These whales were part of the Baltic blue whale population. The specimens are currently housed at the Natural History Museum in Copenhagen and the Geological Museum in Lund. The whales were estimated to be 15-20 years old and were found in the Baltic Sea over 7000 years ago. The bones were dated to 400-900 years old. 

Mark which area you want to distress. VERB

In May, Fujisawa joined Mari Motohashi’s rink as the team’s skip, moving back from Karuizawa to Kitami where she had spent her junior days. 

when are hops added to the brewing process?

The boiling process is where chemical reaction of hops and aromas plus other ingredients create an intense and complex aroma. The boil is extended until the hops are fully extracted. In some small breweries, the boiling process is where chemical reaction of hop flavors, bitterness and aroma compounds through isomerization. Finally, the vapors produced during the boil volatilize off-flavors and intense – a continuous "rolling boil". The boil on average lasts between 60 minutes. At the end of the boil, some of the hops are left behind.

Natural Questions

A Benchmark for Question Answering Research.
Summary

Try out BERT for your tasks

Be aware of what linguistic phenomena are in your data

Try your models on our natural challenge sets

Questions?
Thank you for your attention!

slav@google.com