Learning with Less Data and Labeling for Language Acquisition and Understanding

Elior Sulem

December 2021 - January 2022
Natural Language Processing (NLP)

- **Natural Language Understanding**
  - Extract/deduce information from a text

- **Language Generation**
  - Generate a new text

- **Question-Answering**

- **Dialog**

- **Machine Translation**
Natural Language Processing (NLP)

- In the last few years: New models achieved large improvements on multiple tasks.

- Use of Deep Learning

- Use of Large Pretrained Language Models that dramatically transformed the field

- BERT (Devlin et al., 2019)
- RoBERTa (Liu et al., 2019)
- GPT-2 (Radford et al., 2018)
- GPT-3 (Brown et al., 2020)
In the last few years: New models achieved large improvements on multiple tasks.

“Computers are getting better than humans at reading”,
CNN, January 16, 2018

“It's official, AI is now better at reading comprehension than humans are”,
TechRepublic, January 16, 2018
The Iroquois sent runners to the manor of William Johnson in upstate New York. The British Superintendent for Indian Affairs in the New York region and beyond, Johnson was known to the Iroquois as Warraghiggey, meaning "He who does great things."

What was William Johnson's Iroquois name? Warraghiggey

SQuAD 1.1 (Rajpurkar et al., 2016)
Natural Language Processing (NLP)

- In the last few years: New models achieved large improvements on multiple tasks

![Graph showing F1 score on SQuAD 1.1 (Extractive Question-Answering)]
Current systems do achieve some successes.

My second grader asked me what this passage means: 
"Jupiter is the fifth planet from the Sun and the largest in the Solar System. It is a gas giant with a mass one-thousandth that of the Sun, but two-and-a-half times that of all the other planets in the Solar System combined. Jupiter is one of the brightest objects visible to the naked eye in the night sky, and has been known to ancient civilizations since before recorded history. It is named after the Roman god Jupiter.[19] When viewed from Earth, Jupiter can be bright enough for its reflected light to cast visible shadows,[20] and is on average the third-brightest natural object in the night sky after the Moon and Venus. ""

I rephrased it for him, in plain language a second grader can understand:

Jupiter is the fifth planet from the Sun. It is the largest planet in the Solar System. It is a gas giant. It is very bright, and you can see it at night.

I think he understood it better.
However, they are brittle in realistic situations.

Human: Hello, who are you?
AI: I am an AI created by OpenAI. How can I help you today?

Human: Hi, I'm looking for a restaurant in Tel-Aviv that is vegan and opened on Saturday night. Do you have any suggestions?
AI: I can help with that. Do you prefer Indian or Italian food?

Human: Italian food.
AI: It depends on your price range.

Human: What are the options?
AI: I found restaurants that fit your criteria.

Human: What are their names and addresses?
AI: Do you mean the name of the restaurants or their addresses?

Human: I meant both the names and the addresses of the restaurants
AI: Great question. There are three places that satisfy your requirements, but unfortunately, one restaurant is closed on weekends.
However, they are brittle in realistic situations.

Human: Which restaurants are opened on weekends?
AI: That restaurant, but that one too. I can contact them to see if they let you come on the weekend.
Human: What is their name?
AI: I think you should go to that restaurant because the food is great.

Open-Ended Conversation with an AI Assistant
NLP systems are still far from humans.

A main challenge is **supervision**.

1. Current systems are trained on extremely large amount of data

2. Large-scale annotation is required for every task and domain
Natural Language Processing (NLP)

Natural Language Understanding
Extract/deduce information from a text

Language Generation
Generate a new text

Question-Answering
Dialog
Machine Translation
Pre-training then Fine-tuning Paradigm:

Pre-training on Large Amounts of Unlabeled Data (Billions of Words)

Task-Specific Fine-tuning on Labeled Data

- Question-Answering
- Event Extraction

Examples:
- Project Gutenberg
- Wikipedia: The Free Encyclopedia
Data Challenge

- Pre-training then Fine-tuning Paradigm:

  **Pre-training** on Large Amounts of Unlabeled Data (Billions of Words)

  Task-Specific **Fine-tuning** on Labeled Data

  - Question-Answering
  - Event Extraction
Data Challenge

- **Pre-training on extremely large amounts of unlabeled data**
  - Limits our understanding of low-resource scenarios
  - Infeasible/difficult to apply (training and inference) for many in the academia and industry
Labeling Challenge

Pre-training on Large Amounts of Unlabeled Data (Billions of Words) → Task-Specific Fine-tuning on Labeled Data

- Question-Answering
- Event Extraction
Labeling Challenge

- **Fine-tuning on task-specific labeled data**
  - Progress is limited to specific tasks, in which a lot of annotated data is available.
    - For example, in SQuAD 1.1: 130K examples
  - These models are brittle outside these datasets.
    - The performance usually drops on out-of-domain datasets.

He was arrested for his crimes. When was the arrest? IDK

<table>
<thead>
<tr>
<th>Setting</th>
<th>In-domain</th>
<th>Out-of-domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>80.91</td>
<td>44.78</td>
</tr>
<tr>
<td>Has answer</td>
<td>83.53</td>
<td>68.75</td>
</tr>
<tr>
<td>No answer</td>
<td>78.40</td>
<td>20.80</td>
</tr>
</tbody>
</table>

Performance of a model based on BERT-LARGE and trained on SQuAD 2.0 on in-domain and out-of-domain settings [Sulem et al., 2021]
Labeling Challenge

- Fine-tuning on task-specific labeled data
  - It is not realistic to annotate a lot of data for every task.
  - For information extraction tasks such as event extraction.
    - Usually specific to a particular formalism/ontology.

Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey
Bonan Min*, Hayley Ross*, Elior Sulem*, Amir Pouran Ben Veyseh*, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz and Dan Roth
ArXiv Preprint, 2021
(1) Addressing the Data Challenge

- Pre-training:
  - RoBERTa (Liu et al., 2019) is trained on 30B words.
    - 40 epochs: **1200B words**
  - How many words a 6 years old child has been exposed to?
    - The number of words that a middle-class English-speaking child by the age of 6:
      - no more than **10-50 M** (Hart and Risley, 1995)
    - At that age children have acquired near adult-like grammatical knowledge (Kemp et al., 2005).

<table>
<thead>
<tr>
<th>Model</th>
<th>RoBERTa</th>
<th>Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of words</td>
<td><strong>1200B</strong></td>
<td><strong>10-50M</strong></td>
</tr>
</tbody>
</table>

**BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language**

Philip Huebner, **Elior Sulem**, Cynthia Fisher and Dan Roth
CoNLL 2021, **Best Paper Award Runner Up**
(2) Addressing the Labeling Challenge

- Fine-tuning:
  - Question Answering Formulation:
    - Have large data available and are not specific to a particular ontology
      => QA systems can be probed to solve information extraction problems
  - Some phenomena can be shared across different tasks: unanswerable questions
    - Leveraging other tasks can be useful, in particular in out-of-domain scenarios.

Qing Lyu, Hongming Zhang, Elior Sulem and Dan Roth. ACL 2021

Do We Know What We Don’t Know? Addressing Unanswerable Questions beyond SQuAD 2.0.
Elior Sulem, Jamaal Hay and Dan Roth. EMNLP Findings 2021

Yes, No or IDK. The Challenge of Unanswerable Yes/No Questions.
Elior Sulem, Jamaal Hay and Dan Roth. In Submission
In This Talk

1. Pre-training on Less Data: Language Acquisition Data

2. Fine-tuning with No or Less Task-Specific Labeled Data: QA formulation

3. Research Directions
In This Talk

1. Pre-training on Less Data: Language Acquisition Data

2. Fine-tuning with No or Less Task-Specific Labeled Data: QA formulation

3. Research Directions
Insights from language acquisition in children and its modeling could be useful for improving learning in NLP systems.

On the other hand, Large Pretrained Language Models that led to impressive performance on NLP benchmarks could be good candidates to model language acquisition.

However, current tools do not allow us to make this connection.

- Current models: Children learn from much less words
- Current evaluation (grammaticality tests): Children use a smaller vocabulary
CHILDES dataset

- CHILDES (McWhinney, 2000) includes transcriptions of both child speech and child-directed speech.
- We focus on child-directed speech
- Primarily of in-home recordings of casual speech to children, but also in-lab activities such as book-reading

Examples of child-directed speech forms

**Contraction:**
you wanna go play?

**Dialect differences/grammatical errors:**
is that what you talking about.

**Interruptions and false starts:**
here let’s find ah the gorilla

**Intonation marking:**
That is a real nice building? want me to hold that!

**Made up word forms:**
want to floppity?

**Onomatopeia:**
They go ruff ruff ruff
BERT Pre-training

(Bidirectional Encoder Representation from Transformers)
(Devlin et al., 2019)

**Input tokens**

| CLS | John | went | [MASK] | school | early | SEP | He | was | [MASK] | late | [MASK] | SEP | ... | T512 |

Special symbol in front of the input

Sentence A

Separator
token

Sentence B

Separator
token

John went to school early. He was not late today.

**BERT pre-training** - 2 objectives: Masked LM (MLM) and Next Sentence Prediction (NSP)
BERT Pre-training

(Bidirectional Encoder Representation from Transformers)
(Devlin et al., 2019)

Input tokens:

CLS  John  went  [MASK]  school  early  SEP  He  was  not  late  table  SEP  ...  T512

Special symbol in front of the input

Sentence A

Separator token

Sentence B

Separator token

Unmasking

Random Replacement

Predicts whether B is the actual next sentence of A

80% of the target tokens

Predicts “to”

10% of the target tokens

Predicts “not”

10% of the target tokens

Predicts “today”

BERT pre-training - 2 objectives: Masked LM (MLM) and Next Sentence Prediction (NSP)
RoBERTa Pre-training

80% of the target tokens Predicts “to”

10% of the target tokens Predicts “not”

10% of the target tokens Predicts “today”

• Removing NSP objective
• Bigger batch sizes and longer sentences
• Different masks across epochs

RoBERTa
(Liu et al., 2019)

Input tokens

CLS John went [MASK] school early SEP He was not late table SEP… T512

Special symbol in front of the input

Sentence A
Separator token

Sentence B
Separator token

Unmasking Random Replacement

RoBERTa pre-training - 1 objective: Masked LM (MLM)
From RoBERTa to BabyBERTa (1)

- 6,000x fewer words
- Original RoBERTa: 30B words, Wikipedia and Book-Corpus
- BabyBERTa: 5 M words, child-directed speech transcriptions from CHILDES (McWhinney, 2000)
Compared to pretrained RoBERTa-base, BabyBERTa has:

- **15X fewer parameters**

  - **Size of the model:**
    - Original RoBERTa: 125M Param
      - 12 layers, 12 attention heads, 768 hidden units, intermediate size of 3072
    - BabyBERTa: **8M Param**
      - 8 layers, 8 attention heads, 256 hidden units, intermediate size of 1024

- **6X smaller vocabulary**

  - **Vocabulary size:**
    - Original RoBERTa: 50265
    - BabyBERTa: **8192**
BabyBERTa Pre-training

- **BabyBERTa Unmasking**
  - Probability = 0 (No unmasking)
  - Masks force the model to attend to lexical context in order to make predictions.

90% of the target tokens
- Predicts “to”

10% of the target tokens
- Predicts “not”
- Predicts “today”

Input tokens:

<table>
<thead>
<tr>
<th>CLS</th>
<th>John</th>
<th>went</th>
<th>[MASK]</th>
<th>school</th>
<th>early</th>
<th>SEP</th>
<th>He</th>
<th>was</th>
<th>[MASK]</th>
<th>late</th>
<th>table</th>
<th>SEP</th>
<th>T512</th>
</tr>
</thead>
</table>

Special symbol in front of the input

Sentence A

Separator token

Sentence B

Separator token

* Removing Random Replacement does not affect our results
A new model: BabyBERTa

- Based on RoBERTa (Liu et al., 2019)
- Training data:
  - Original RoBERTa: 30B words, Wikipedia and Book-Corpus (Zhu et al., 2015)
  - BabyBERTa: 5 M words, child-directed speech transcriptions from CHILDES (McWhinney, 2000)
- Size of the model:
  - Original RoBERTa: 12 layers, 12 attention heads, 768 hidden units, intermediate size of 3072
  - BabyBERTa: 8 layers, 8 attention heads, 256 hidden units, intermediate size of 1024
- Vocabulary size:
  - Original RoBERTa: 50265
  - BabyBERTa: 8192
- Unmasking Probability:
  - Original RoBERTa: 0.10
  - BabyBERTa: 0 (No unmasking)
    - Masks force the model to attend to lexical context in order to make predictions.
One of the ways to probe language representations is to test on specialized datasets addressing a specific phenomenon. (e.g. Linzen et al., 2016; Goldberg, 2019)

BLiMP dataset (Warstadt et al., 2020)
- 12 grammatical phenomena
- 67 small datasets
- 1,000 minimal pairs in each dataset
- Isolate specific phenomena in syntax, morphology, or semantics.
Probing via Grammaticality Tests

- **Example 1: Noun-Verb Agreement (from BLiMP)**
  - **Acceptable example:** These casseroles *disgust* Kayla.
  - **Unacceptable example:** These casseroles *disgusts* Kayla.

- **Example 2: Irregular Verbs (from BLiMP)**
  - **Acceptable example:** Aaron *broke* the unicycle.
  - **Unacceptable example:** Aaron *broken* the unicycle.

- The test sentences in BLiMP are not adapted to the CHILDES vocabulary.
A New Grammar Test Suite

- Adapted to the CHILDES vocabulary

- Lists of words (nouns, adjectives, verbs) counterbalanced to compare between three corpora:
  - CHILDES,
  - Newsela (Xu et al., 2015; simplified text)
  - Wikipedia

- New Grammar Test Suite
  - 13 grammatical phenomena
  - 23 paradigms
  - 2,000 minimal pairs for each paradigm
Does BabyBERTa “know” grammar?

- Experiments:
  - New Grammar Test Suite
  - Preference Score (when comparing the two sentences of the minimal pair):
    - Summing the cross-entropy errors at each position in the sentence (Zaczynska et al., 2020)
    - Accuracy: dividing the number of correct choices by the total number of pairs.

- Results:

<table>
<thead>
<tr>
<th>Model (Data Size)</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa-base – Liu et al., 2019 (30B)</td>
<td>81.1</td>
</tr>
<tr>
<td>RoBERTa-base - Warstadt et al., 2020 (10M)</td>
<td>64.5</td>
</tr>
<tr>
<td>RoBERTa-base on CHILDES (5M)</td>
<td>59.2</td>
</tr>
<tr>
<td>BabyBERTa with unmasking (5M)</td>
<td>56.4</td>
</tr>
<tr>
<td>BabyBERTa (5M)</td>
<td><strong>80.5</strong></td>
</tr>
</tbody>
</table>

Average accuracy on our grammar test suite
Does BabyBERTa “know” grammar?

- Compared to pretrained RoBERTa-base, BabyBERTa has:
  - 15X fewer parameters
  - 6X smaller vocabulary
Does BabyBERTa “know” grammar?

問い: 6,000X fewer words
Does BabyBERTa “know” grammar?

- However, BabyBERTa performs comparably to pre-trained RoBERTa-base.

<table>
<thead>
<tr>
<th></th>
<th>RoBERTa-base</th>
<th>BabyBERTa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware (GPU)</td>
<td>1024x V100</td>
<td>1x GTX1080</td>
</tr>
<tr>
<td>Training Time</td>
<td>24 hours</td>
<td>2 hours</td>
</tr>
<tr>
<td>Average Accuracy</td>
<td>81.0</td>
<td>80.5</td>
</tr>
</tbody>
</table>

Comparison between RoBERTa base and BabyBERTa, including the average accuracy on our grammar test suite.
Is it specific to CHILDES?

- We replace CHILDES by data from Newsela and Wikipedia with the same number of sentences.
- BabyBERTa trained on Wikipedia performs well below the others on paradigms involving questions. Indeed, questions correspond to 40% of our CHILDES corpus and no more than 1% in Wikipedia.
- Overall, Newsela (compiled for pedagogical purpose) and CHILDES achieve better results than Wikipedia.

### Comparison between BabyBERTa trained on CHILDES, Newsela and Wikipedia on our grammar test suite

<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHILDES</td>
<td>77.2</td>
</tr>
<tr>
<td>Newsela</td>
<td><strong>79.0</strong></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>73.0</td>
</tr>
</tbody>
</table>

Newsela: English news articles, and 4 or 5 simplified versions of each rewritten by professional annotators for children with different reading proficiency.

The models are trained on during the same number of steps for each condition.
Is CHILDES a good starting point for training?

- We compare the order of training in two experiments, keeping the order of appearance of the sentences in each corpus.

| 1) | CHILDES + Newsela | Average Accuracy | 80.3 |
| 2) | Wikipedia + Newsela | 77.8 |
| 2) | Newsela + Wikipedia | 78.4 |

Suggests that CHILDES is a good starting point.

- Newsela is not necessarily a better end point.

Not statistically different*
Summary

- New tools: (i) a new model and a (ii) new grammar test for the use of Pretrained Language Models for modeling language acquisition.

- Investigating learning-related questions relevant to both language acquisition modeling and NLP:
  - LMs can achieve good performance on grammaticality tests with inputs available to children:
    - 5 M instead of 30B
  - The domain is important: Newsela > CHILDES > Wikipedia
  - CHILDES is a good starting point for training, at least when less data is available.
Future Directions

- CHILDES also differs from Wikipedia in targeting speech rather than written language.
  - Experimenting with transcriptions of Adult Spoken Language.

- Unmasking may be important for downstream tasks.
  - Exploring the best alternation between masking and unmasking over time.
  - Experimenting on downstream tasks

- We focus on language information available to children.
  - Interactions with other modalities such as sound and vision (Goodman et al., 2007)
In This Talk

1. Pre-training on Less Data: Language Acquisition Data

2. Fine-tuning with No or Less Task-Specific Labeled Data: QA formulation

3. Research Directions

Learning with Less Data and Labeling

Deep learning Model

https://www.freevector.com
The US Centers for Disease Control and Prevention is recommending people wear face coverings in public and health officials just reported the most deaths in a single day. CNN – April 3 2020

The European Union’s health and aviation bodies have issued a new set of guidelines for air travel, recommending the use of face masks and the practice “scrupulous and frequent” hand hygiene on flights in order to ensure safety of travelers and aviation personnel amid the Covid-19 pandemic. CNN – May 20 2020

Face coverings will become mandatory again in shops and on public transport in England from next week as part of measures to target the new coronavirus variant, Omicron, the PM has said. BBC – November 28 2021

On May 13 [2020], the Centers for Disease Control and Prevention said that Americans who are fully vaccinated against the coronavirus may stop wearing masks or maintaining social distance in most indoor and outdoor settings, regardless of size. NYT – April 27 2021
The US Centers for Disease Control and Prevention is recommending people wear face coverings in public and health officials just reported the most deaths in a single day. CNN – April 3 2020

The European Union’s health and aviation bodies have issued a new set of guidelines for air travel, recommending the use of face masks and the practice “scrupulous and frequent” hand hygiene on flights in order to ensure safety of travelers and aviation personnel amid the Covid-19 pandemic. CNN – May 20 2020

Face coverings will become mandatory again in shops and on public transport in England from next week as part of measures to target the new coronavirus variant, Omicron, the PM has said. BBC– November 28 2021

On May 13 [2020], the Centers for Disease Control and Prevention said that Americans who are fully vaccinated against the coronavirus may stop wearing masks or maintaining social distance in most indoor and outdoor settings, regardless of size. NYT – April 27 2021
There is a large amount of unstructured text

Event extraction allows us to provide structures:
- Different types of events: recommendations, regulations, contamination, ...
- Different locations: countries, states, cities, ...
- Different times and dates
- Different participants: CDC, The European Union’s health and aviation bodies, governments

The information can be then situated according to the different dimensions (Dror et al., 2021)
Event Extraction

- **Input:** “China purchased two nuclear submarines from Russia last month.”

- **Output:** Event type: TRANSFER-OWNERSHIP

  China has purchased two nuclear submarines from Russia last month.

  - **Buyer-Arg:** China
  - **Trigger:** purchased
  - **Artifact-Arg:** two nuclear submarines
  - **Seller-Arg:** Russia
  - **Time-Arg:** last month

- **Subtasks:** Trigger Identification (TI), Trigger Classification (TC), Argument Identification (AI), Argument Classification (AC).
Most current work on Events is based on supervised learning.

Large amounts of text have been annotated at a rather deep level.

- Costly, requires expertise, leads to inconsistencies (across, and even within, datasets)
- Limited to specific domains and a limited event ontology.
- What if we want to identify new types of events and their structure (arguments)?
Main Thesis: When an event schema library is given

- Definitions of events of interest

Then extracting an event expressed in text reduces to answering a small number of schema-driven questions about the text.

This gives rise to transferring event extraction capabilities from QA-supporting models, without task-specific training on event datasets.

- The same QA model can be applied to different Event Datasets and Domains (e.g. financial, medical).

Qing Lyu, Hongming Zhang, Elior Sulem and Dan Roth. ACL 2021
Event Extraction as Question Answering

- **Input:** China purchased two nuclear submarines from Russia last month.

- **Trigger:** purchased

- **Event Type:**
  - Q0: Did someone transfer ownership?  
    A0: Yes \(\Rightarrow\) TRANSFER-OWNERSHIP (TC)  
  - (multiple questions are being asked)

- **Arguments:** (now we know the event type)
  - Q1: What was purchased?  
    A1: Two nuclear submarines. \(\Rightarrow\) Artifact-Arg  
  - (multiple questions for each arg type)

  - Q2: Who purchased two nuclear submarines?  
    A2: China. \(\Rightarrow\) Buyer-Arg

  - Q3: Who did China purchase two nuclear submarines from?  
    A3: Russia. \(\Rightarrow\) Seller-Arg
Experiments on the ACE dataset in a zero-shot approach

- We propose the first zero-shot approach based on transfer learning for both triggers and arguments.
  - **Works with any ontology** – even when no training data exists.
  - **Argument extraction** is formalized as a schema-driven sequence of probing questions
  - **QA model**: RoBERTa (Liu et al., 2019) trained on QAMR (Michael et al., 2018).

### Zero-shot Approaches

<table>
<thead>
<tr>
<th>Setting</th>
<th>Previous</th>
<th>SOTA</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>supervised</td>
<td>56.8</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>scratch</td>
<td>15.8</td>
<td>16.8</td>
<td></td>
</tr>
<tr>
<td>Gold TI</td>
<td>14.7</td>
<td>24.2</td>
<td></td>
</tr>
<tr>
<td>Gold TI+TC</td>
<td>25.8</td>
<td>27.4</td>
<td></td>
</tr>
</tbody>
</table>

**F1 score for Argument Identification + Argument Classification on ACE 2005**

**Improvement relative to previous unsupervised approaches but still a large gap compared to supervised methods**
Not all the possible arguments in the schema will appear in a given sentence.

**Input:** *China purchased two nuclear submarines.*

When did China purchase two nuclear submarines? **No answer**
What was William Johnson's Iroquois name?

Warraghiggey

SQuAD 1.1 (Rajpurkar et al., 2016)
The Iroquois sent runners to the manor of William Johnson in upstate New York. The British Superintendent for Indian Affairs in the New York region and beyond, Johnson was known to the Iroquois as Warraghiggey, meaning "He who does great things."

What was William Johnson's Sioux name?
The Iroquois sent runners to the manor of William Johnson in upstate New York. The British Superintendent for Indian Affairs in the New York region and beyond, Johnson was known to the Iroquois as Warraghiggey, meaning "He who does great things."

What was William Johnson's Sioux name?
I don't know

SQuAD 2.0 (Rajpurkar et al., 2018)

<table>
<thead>
<tr>
<th></th>
<th>train test</th>
<th>SQuAD 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td></td>
<td>80.91</td>
</tr>
<tr>
<td>Has answer</td>
<td></td>
<td>83.53</td>
</tr>
<tr>
<td>No answer</td>
<td></td>
<td>78.40</td>
</tr>
</tbody>
</table>

In-domain dev performance (F1) for a BERT-LARGE model fine-tuned on SQuAD 2.0.
Informative evaluation requires out-of-domain test sets
- controlled out-of-domain test sets (Linzen, 2020)
- Ask very simple questions whose answer is obvious to humans. (Dunietz et al. 2020)

QA applications involve out-of-domain test sets
- Zero-shot event extraction (Lyu et al., 2021)
- Evaluation of summarization (Deutsch et al. 2021)
New Event-Based Test Dataset

- Compiling in semi-automatic way a test event corpus for wh-questions - **ACE-whQA**, derived from ACE, focusing on time and location: 734 examples
  - **Has-answer:**
    - She lost her seat in the 1997 election.
    - When was the loss?
  - **Competitive IDK:**
    - She travelled to **Mexico** after she lost her seat in the 1997 election
    - Where was the loss?
  - **Non-Competitive IDK:**
    - He was arrested for his crimes
    - When was the arrest?
Leveraging Textual Entailment

- RTE task (Dagan et al., 2013)
  - Given a Premise and a Hypothesis: 3 labels: Entailment/Contradiction/Neutral

- MNLI dataset (Williams et al., 2018)

<table>
<thead>
<tr>
<th>Premise:</th>
<th>John was born in New York.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis:</td>
<td>John was born in France.</td>
</tr>
<tr>
<td>Label:</td>
<td>Contradiction</td>
</tr>
</tbody>
</table>
## Out-of-domain Performance

### Evaluation on ACE-whQA:

- Low performance of a top system trained on SQuAD 2.0 (Rajpurkar et al., 2018)
- First training on Textual Entailment (Dagan et al., 2013) that includes an IDK option (“neutral”) improves the performance, in particular for non-competitive IDK questions.

### Baseline vs. Using TE

<table>
<thead>
<tr>
<th></th>
<th>train</th>
<th>test</th>
<th>SQuAD 2.0</th>
<th>MNLI + SQuAD 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has Answer</td>
<td></td>
<td></td>
<td>68.75</td>
<td>71.68</td>
</tr>
<tr>
<td>Compet. IDK</td>
<td></td>
<td></td>
<td>20.80</td>
<td>46.40*</td>
</tr>
<tr>
<td>Non-Compet. IDK</td>
<td></td>
<td></td>
<td>28.46</td>
<td>75.61*</td>
</tr>
</tbody>
</table>

F1 scores of the BERT-LARGE system evaluated on ACE-whQA.

* Significantly higher than the baseline (p<0.05)
IDK Questions

- We found that leveraging RTE is particularly useful for out-of-domain test sets with IDK in Extractive QA.

- SQuAD 2.0 alone is not sufficient to address IDK in these cases.
IDK Questions - Future Directions

- Addressing IDK in adversarial cases, namely the competitive IDK.

- Extending the study of IDK to additional tasks.

Human: Hello, who are you?
AI: I am an AI created by OpenAI.
How can I help you today?
Human: Hi, do you know who is the President of Spain?
AI: The President of Spain is Mr. Mariano Rajoy, who was elected on 20 December 2011 by the Congress of Deputies, replacing José Luis Rodríguez Zapatero who had been president since 2004. Mr. Rajoy's party also won a plurality, with 186 seats in the Senate.

Open-Ended Conversation with an AI Assistant
Conclusion

- Natural Language Understanding requires the training of models.
- Current systems achieve limited success, but they are not scalable.
- We address it from multiple perspectives:
  - Realistic training of Language Models: **Less Data**
  - New formulations that facilitate zero-shot Information Extraction: **Less Labeling**
Future Directions

Natural Language Understanding

- Not necessarily explicit
- In multiple languages
- In a scalable way

- Child Language Acquisition
- Prompt-based Learning
Future Directions: Implicit Information Extraction

- **Extract Implicit Relations**

  **Sentence:** John gave Jane the book he got from Bill.
  **Question:** Who has the book?

  ![Diagram showing relationships between Bill, John, and Jane]

  **Sentence:** John described the situation in a letter to Bill.
  **Question:** Who wrote to Bill?
Future Directions: Implicit Information Extraction

- Current work: Data collection using crowdsourcing to evaluate the ability of current models to detect implicit information. What have we learned about this participant?

- How can we improve generalization?

- Commonsense and Grounding Information

- Supervision challenges
Future Directions: Multilingual Understanding

- Large corpora are missing in many languages
- Transfer Across languages
- Multilingual Representations: mBERT (Devlin et al., 2019), XLMR (Conneau et al., 2019)
  - Cross-lingual Event Extraction using Question Answering
    Work in process with Tianyi Zhang, Yee Seng, Kemanth Kandula, Bonan Min and Dan Roth.
Future Directions: Multilingual Understanding

- Using symbolic representations that are stable across languages

**Universal Conceptual Cognitive Annotation** (Abend and Rappoport, 2013)

- Scenes are evoked by a **Main Relation** (Process/State)
- A Scene may contain one or several **Participants**.

Parallel Scene (H)  Linker (L)  Participant (A)  Process (P)

He came back home played piano

He and she played piano at home.

*He* came back from playing the piano.
Future Directions: Multilingual Understanding

- Within the same language, consistency across paraphrases
- Also combining multiple modalities: Visual Question Answering
- Visual information as a bridge across languages
Future Directions: Prompt-based Learning

- Prompt-based learning
  - Auxiliary Tasks: Question-Answering and Textual Entailment [In this Talk]
  - Template-based Prompts: make Pre-training and Fine-tuning similar

  **Example for Textual Entailment** (Schick and Schutze, 2021)
  Mia likes pie? _____, Mia hates pie!
  No $\rightarrow$ contradiction.

- Relation between Pre-training and Fine-tuning
  - Predicate argument structure - Work in progress with Chaitanya Malaviya, Xingyu Fu, Mark Yatskar, Charles Yang and Dan Roth
How can we learn natural language in an efficient way?
Do we need explicit linguistic knowledge and structures to learn natural language?

These questions are asked both in Psycholinguistics, to explore the way children learn language, and in Natural Language Processing (NLP), to build efficient systems that operate on natural language.
Future Directions: Language Acquisition

- Modeling language acquisition using NLP models
- Using Insights from language acquisition modeling to build better systems
- Computational models of bilingualism and multilingualism.
- Taking into account additional modalities: images, videos, sounds. (e.g. Kádár et al., 2019).
Future Directions: Language Acquisition

- Leveraging the questions in CHILDES to create a QA dataset
- Pre-training on QA data (He et al., 2020, Chen et al., 2020)
Additional Works

- Conceptual Annotations Preserve Structure Across Translations
  Elior Sulem, Omri Abend and Ari Rappoport, S2MT 2015

- Semantic Structural Decomposition for Neural Machine Translation
  Elior Sulem, Omri Abend and Ari Rappoport, *SEM 2020

- Simple and Effective Text Simplification Using Semantic and Neural Methods
  Elior Sulem, Omri Abend and Ari Rappoport, ACL 2018

- Semantic Structural Evaluation for Text Simplification
  Elior Sulem, Omri Abend and Ari Rappoport, NAACL 2018

- BLEU is not Suitable for Evaluation of Text Simplification
  Elior Sulem, Omri Abend and Ari Rappoport, EMNLP 2018

- The Language of Legal and Illegal Activity in the Darknet.
  Leshem Choshen*, Dan Eldad*, Daniel Hershcovich*, Elior Sulem*, Omri Abend, ACL 2019

- Capturing the Content of a Document through Complex Event Identification.
  Zheng Qi, Elior Sulem, Haoyu Wang, Xiaodong Yu and Dan Roth, In submission.
Research Collaborators

Penn

Jamaal Hay
Hongming Zhang
Xiaodong Yu
Helen Jin
Tianyi Zhang
Zheng Qi

Prof. Dan Roth
Qing Lyu
Haoyu Wang
Dr. Eleni Miltsakaki
Prof. Mark Yatskar
Prof. Charles Yang
Chaitanya Malaviya
Xingyu Fu

UNIVERSITY OF ILLINOIS
URBANA-CHAMPAIGN

Philip Huebner
Prof. Cindy Fisher
Amir Pouran Ben Veyseh,
Prof. Thien Huu Nguyen

Prof. Eneko Agirre
Oscar Seinz

Prof. Eniko Agirre

Hayley Ross

Raytheon

BBN

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Dr. Bonan Min
Ilana Heinz
Yee Seng
Kemanth Kandula

THE HEBREW UNIVERSITY OF JERUSALEM

Dan Eldad
Leshem Choshen
Dr. Daniel Hershcovich

Prof. Ari Rappoport

Prof. Omri Abend
Thank you

Pre-training on Less Data: Language Acquisition Data

Fine-tuning with No or Less Task-Specific Labeled Data: QA formulation

Deep learning Model

https://www.freevector.com

eliors@seas.upenn.edu
https://www.cis.upenn.edu/~eliors/