Natural Language Understanding and Inference: Benchmarks, Resources, and Approaches

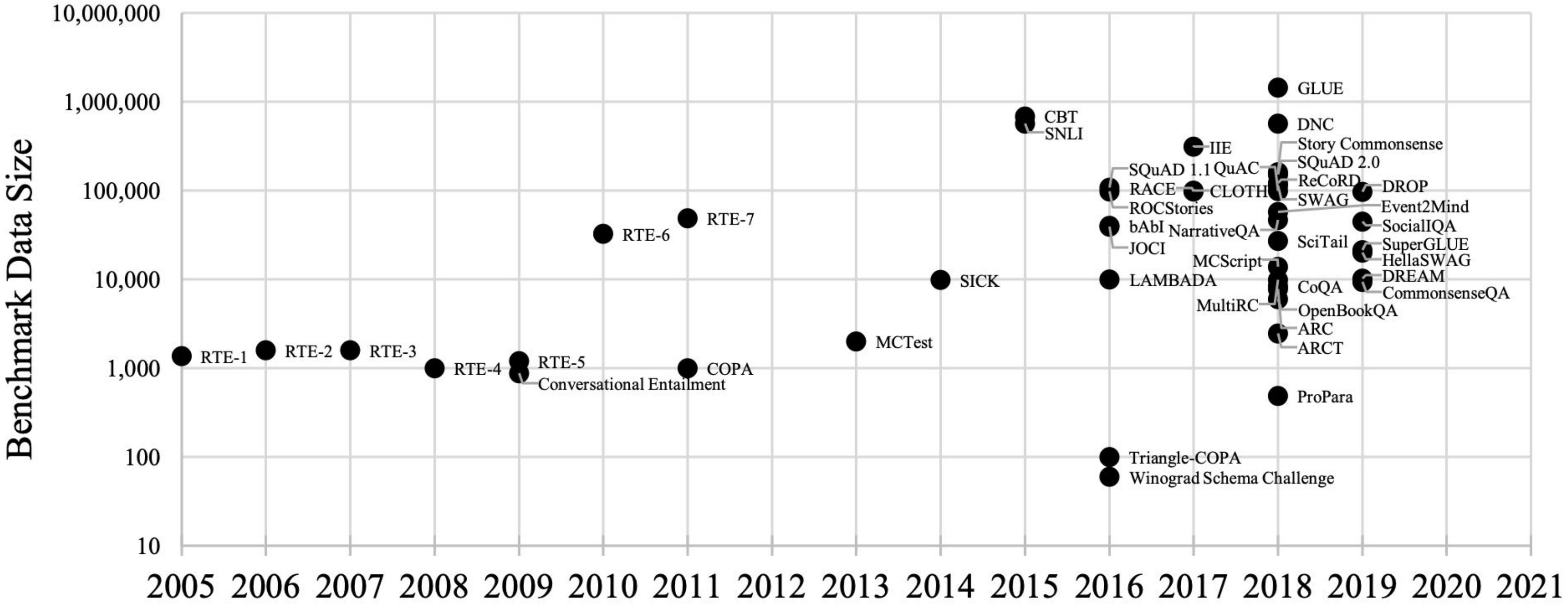
Shane Storks (University of Michigan)
Qiaozi Gao (Michigan State University)
Joyce Y. Chai (University of Michigan)

Understanding Natural Language

"Jack needed some money, so he went and shook his piggy bank.
He was disappointed when it made no sound."
- Why was Jack disappointed? (Minsky, 2000)

- **Benchmarks** that require deep language understanding that goes beyond what's explicitly written, and rely on inference and knowledge of the world.
- **Knowledge**
  - linguistic knowledge (e.g., Penn Treebank, WordNet)
  - common knowledge (e.g., Freebase, DBpedia, YAGO)
  - commonsense knowledge (e.g., ConceptNet, ATOMIC)

Benchmarks: Data Size



Benchmarks

- Coreference Resolution
  - e.g., Winograd Schema Challenge
- Question Answering
  - e.g., SQuAD, OpenBookQA
- Textual Entailment
  - e.g., RTE, SNLI
- Plausible Inference
  - e.g., COPA, ROCStories
- Multiple Tasks
  - e.g., GLUE, DNC

Benchmarks

- **Coreference Resolution**
  - **e.g., Winograd Schema Challenge**
- Question Answering
  - e.g., SQuAD, OpenBookQA
- Textual Entailment
  - e.g., RTE, SNLI
- Plausible Inference
  - e.g., COPA, ROCStories
- Multiple Tasks
  - e.g., GLUE, DNC

- The trophy would not fit in the brown suitcase because it was too **big**.
- What was too **big**?

A. The trophy
B. The suitcase

Benchmarks

- **Coreference Resolution**
  - **e.g., Winograd Schema Challenge**
- Question Answering
  - e.g., SQuAD, OpenBookQA
- Textual Entailment
  - e.g., RTE, SNLI
- Plausible Inference
  - e.g., COPA, ROCStories
- Multiple Tasks
  - e.g., GLUE, DNC

- The trophy would not fit in the brown suitcase because it was too **small**.
- What was too **small**?

A. The trophy
B. The suitcase

Benchmarks

- Coreference Resolution
  - e.g., Winograd Schema Challenge
- **Question Answering**
  - **e.g., SQuAD, OpenBookQA**
- Textual Entailment
  - e.g., RTE, SNLI
- Plausible Inference
  - e.g., COPA, ROCStories
- Multiple Tasks
  - e.g., GLUE, DNC

- Which of these would let the most heat travel through?

A. a new pair of jeans.
B. a steel spoon in a cafeteria.
C. a cotton candy at a store.
D. a calvin klein cotton hat.

Evidence: Metal is a thermal conductor.

Benchmarks

- Coreference Resolution
  - e.g., Winograd Schema Challenge
- Question Answering
  - e.g., SQuAD, OpenBookQA
- **Textual Entailment**
  - **e.g., RTE, SNLI**
- Plausible Inference
  - e.g., COPA, ROCStories
- Multiple Tasks
  - e.g., GLUE, DNC

- ***Text:*** A black race car starts up in front of a crowd of people.
- ***Hypothesis:*** A man is driving down a lonely road.
- ***Label:*** contradiction

Benchmarks

- Coreference Resolution
  - e.g., Winograd Schema Challenge
- Question Answering
  - e.g., SQuAD, OpenBookQA
- Textual Entailment
  - e.g., RTE, SNLI
- **Plausible Inference**
  - **e.g., COPA, ROCStories**
- Multiple Tasks
  - e.g., GLUE, DNC

I knocked on my neighbor's door.
What happened as result?

A. My neighbor invited me in.
B. My neighbor left his house.

Benchmarks

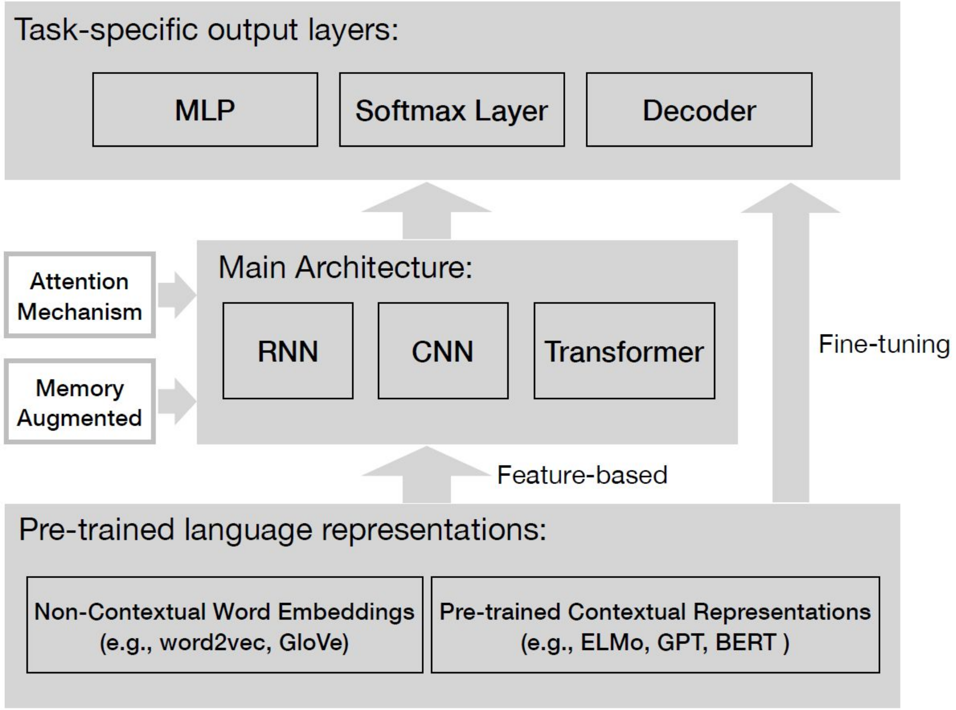
- Coreference Resolution
  - e.g., Winograd Schema Challenge
- Question Answering
  - e.g., SQuAD, OpenBookQA
- Textual Entailment
  - e.g., RTE, SNLI
- Plausible Inference
  - e.g., COPA, ROCStories
- **Multiple Tasks**
  - **e.g., GLUE, DNC**

Creating Benchmarks: Criteria and Considerations

- Task Format
  - Classification tasks
  - Open-ended tasks
- Evaluation Scheme
  - Evaluation metrics: objective and easy to calculate
  - Human performance measurement
- Avoiding Data Biases
  - Label distribution bias
  - Question Type Bias in QA
  - Superficial Correlation Bias (gender bias, human stylistic artifacts)

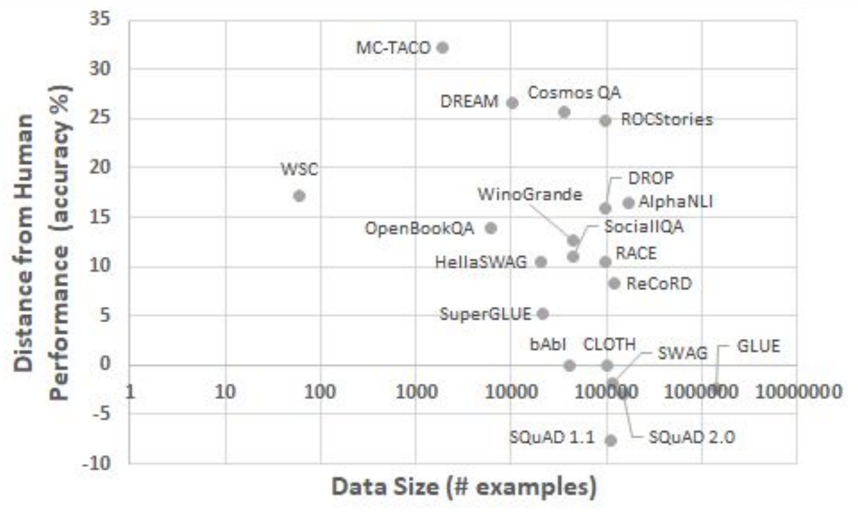
Approaches: General Architecture

- Symbolic approaches
- Statistical approaches
- Latest SOTA use deep neural network (e.g., transformer) with built-in pre-trained contextual embeddings
  - Performance keeps increasing
  - Exceeding human performance sometimes



Performance Trends

- Many factors may affect progress on benchmarks
  - Actual task difficulty
  - Data size
  - Year released
  - Number of people working on the benchmark
  - Data bias
- Performance should be interpreted with caution



Future Questions

- Doe the benchmark performance really reflect the machine inference abilities?
- How to explain model behaviors so that humans can understand the underlying inference process?
- How can we make better use of available knowledge resources?
- How can we train energy/cost efficient models?
  - How the Transformers broke NLP leaderboards - Rogers, 2019
  - Green AI - Schwartz et al., 2019

Creating Benchmarks: Data Biases

- Label Distribution Bias
  - relatively easy to avoid: an equal number of examples for each class
- Question Type Bias in QA
  - distribution of the first words of questions (e.g., CoQA, CommonsenseQA)
  - manually analysis of question categories (e.g., Squad 2.0, ARC)
  - predefined question types (e.g., ProPara)
- Superficial Correlation Bias
  - e.g., gender bias, human stylistic artifacts
  - relatively difficult to avoid
  - adversarial filtering process (e.g., SWAG)

Benchmarks

- Turing Test
  - encouraging machines to deceive humans
  - no feedback on a continuous scale to allow for incremental development

- Early NLP Benchmarks
  - Part-of-speech Tagging
  - Named Entity Recognition
  - Coreference Resolution
  - Information Extraction

Jyc: delete this slide

Jyc: at least show two or three slides about approaches:

- One slide on the general architecture
- One slide on example performance? Shane is making a figure for that, discuss the differences between human performance and model performance.

Thank you!

Also need a slide to summarize:

- What pending questions from the exercise on benchmarks.
- What should be some ideas for future direction.

Knowledge Base

Humans perform inference based on vast amount of knowledge about how the world works. To support machines' inference ability, a parallel ongoing research effort in the last several decades is the development of various knowledge resources.

Knowledge Base Collection

Discuss issues related to collecting knowledge required to perform commonsense reasoning

Learning and Inference Approaches

- Symbolic Approaches
- Statistical Approaches
- Neural Approaches

Model Generalization

Consequence of previous issue?

Talk about current SOTA models and probing studies (like Niven and Kao, 2019)