Recent Advances in Natural Language Inference

许亮 51184501169

Introduction

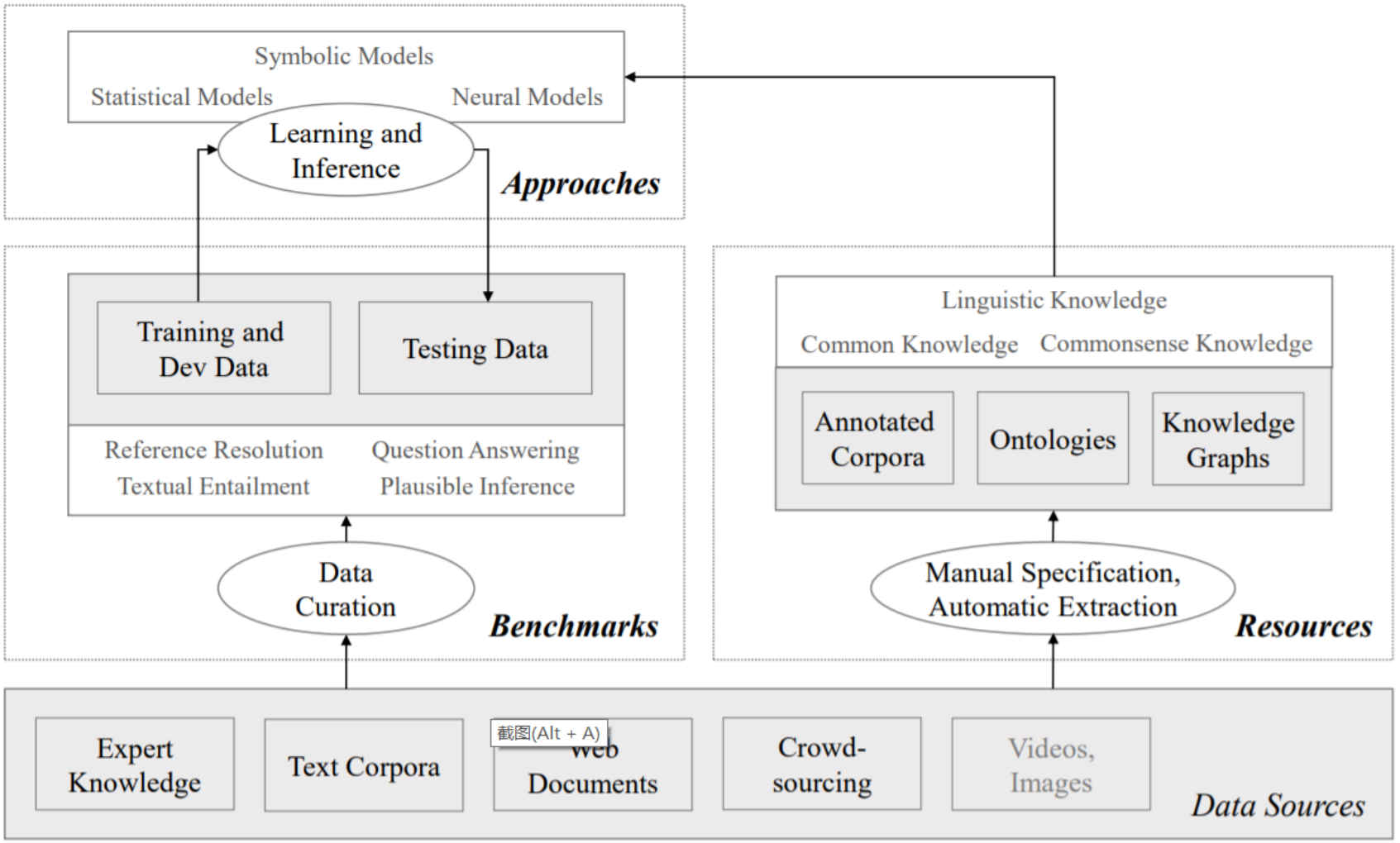
- In the NLP community, recent years have seen a surge of research activities that address machines' ability to perform deep language understanding which goes beyond what is explicitly stated in text, rather relying on reasoning and knowledge of the world.

- While this kind of knowledge and reasoning comes so naturally to human readers, it is notoriously difficult for machines. Despite significant advances in natural language processing in the last several decades, machines are still far away from having this type of natural language inference (NLI) ability

- To address this problem, recent years have seen a surge of research activities on NLI: machines' capability of deep understanding of language that goes beyond what is explicitly expressed, rather relying on new conclusions inferred from knowledge about how the world works.

Three areas

- Benchmark
- Resources
- Approaches


Symbolic Models
Statistical Models
Neural Models
Learning and Inference
Approaches
Training and Dev Data
Testing Data
Reference Resolution
Textual Entailment
Question Answering
Plausible Inference
Data Curation
Benchmarks
Linguistic Knowledge
Common Knowledge
Commonsense Knowledge
Annotated Corpora
Ontologies
Knowledge Graphs
Manual Specification, Automatic Extraction
Resources
Expert Knowledge
Text Corpora
Web Documents
Crowd-sourcing
Videos, Images
Data Sources

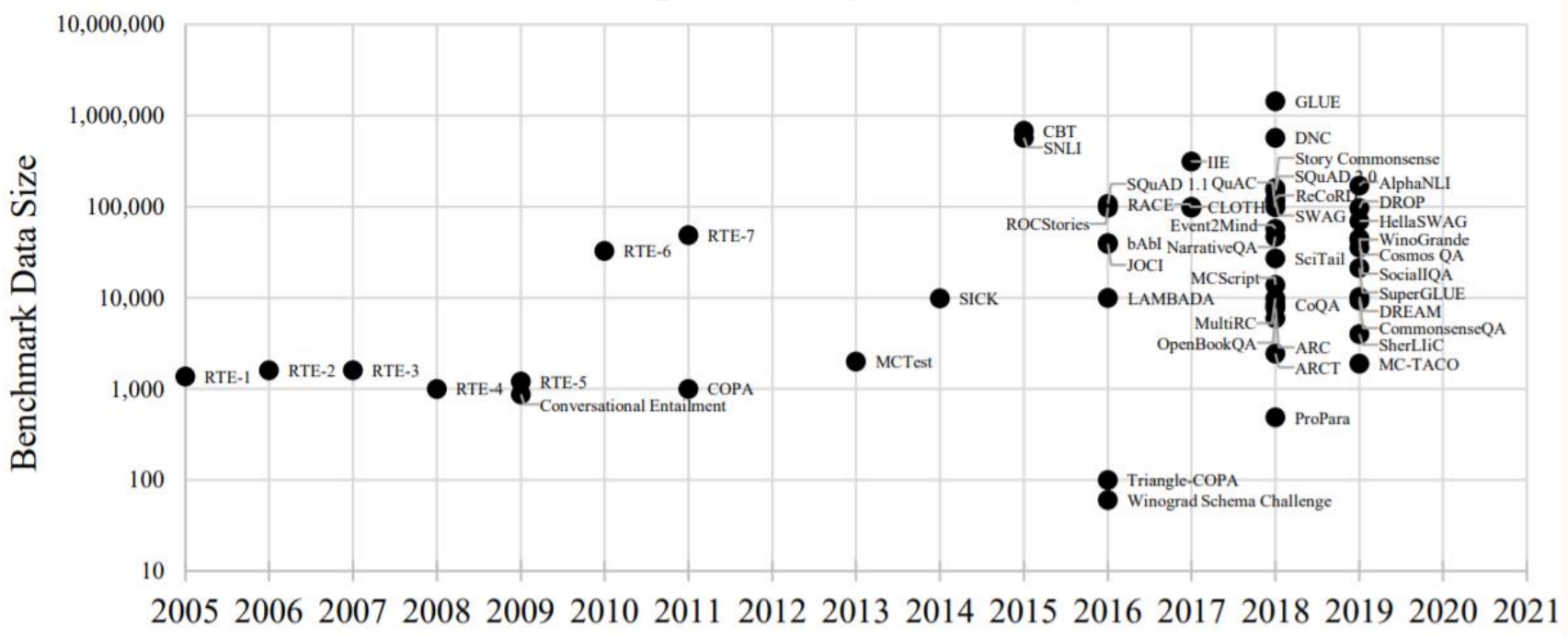
Scope of Knowledge Resources

- Linguistic knowledge
- Common knowledge
- Commonsense knowledge

Learning and Inference Approaches

- Symbolic Approaches
- Early Statistical Approaches
- Neural Approaches
- Incorporating External Knowledge

Benchmarks and Tasks



REFERENCE RESOLUTION

(A) Winograd Schema Challenge (Levesque, 2011)

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

a. **The trophy**
b. The suitcase

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

a. The trophy
b. **The suitcase**

(B) Winogender (Rudinger et al., 2018a)

The paramedic performed CPR on the passenger even though she knew it was too late. Who knew it was too late?

a. **The paramedic**
b. The passenger

(C) (Rahman & Ng, 2012)

Lions eat zebras because they are predators. Who are predators?

a. **Lions**
b. Zebras

QUESTION ANSWERING

- Question answering (QA) is one such comprehensive task, particularly the recent formulation of the task providing a passage, and requiring a system to answer questions about it to demonstrate its comprehension of the passage.

- QA is a fairly well-established area in NLP, and there are many existing benchmarks for QA:
  - CoQA
  - SQuAD
  - OpenBookQA
  - RACE
  - NarrativeQA
  - MultiRC
  - CommonsenseQA
  - DREAM

(A) MCScript (Ostermann et al., 2018)
Did they throw away the old diaper?

a. **Yes, they put it into the bin.**
b. No, they kept it for a while.

(B) OpenBookQA (Mihaylov et al., 2018)
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.

(C) CoQA (Reddy et al., 2018)
The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Who is the democratic candidate?
Terry McAuliffe
Evidence: Democrat Terry McAuliffe

Who is his opponent?
Ken Cuccinelli
Evidence: Republican Ken Cuccinelli

TEXTUAL ENTAILMENT

- Textual entailment is defined by Dagan et al. (2005) as a directional relationship between a text and a hypothesis, where it can be said that the text entails the hypothesis if a typical person would infer that the hypothesis is true given the text. Some benchmarks expand this task by also requiring recognition of contradiction.

- several textual entailment :
  - Recognizing Textual Entailment (RTE)
  - Conversational Entailment
  - Sentences Involving Compositional Knowledge (SICK)
  - Stanford Natural Language Inference (SNLI)
  - SciTail
  - SherLIiC

(A) RTE Challenge (Dagan et al., 2005)

Text: American Airlines began laying off hundreds of flight attendants on Tuesday, after a federal judge turned aside a union's bid to block the job losses.

Hypothesis: American Airlines will recall hundreds of flight attendants as it steps up the number of flights it operates.

Label: **not entailment**

(B) SICK (Marelli, Menini, Baroni, Bentivogli, Bernardi, & Zamparelli, 2014a)[5]

Sentence 1: Two children are lying in the snow and are drawing angels.

Sentence 2: Two children are lying in the snow and are making snow angels.

Label: **entailment**

(C) SNLI (Bowman et al., 2015)

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**

(D) MultiNLI, Telephone (Williams, Nangia, & Bowman, 2017)

Context: that doesn't seem fair does it

Hypothesis: There's no doubt that it's fair.

Label: **contradiction**

(E) SciTail (Khot, Sabharwal, & Clark, 2018)

Premise: During periods of drought, trees died and prairie plants took over previously forested regions.

Hypothesis: Because trees add water vapor to air, cutting down forests leads to longer periods of drought.

Label: **neutral**

PLAUSIBLE INFERENCE

(A) COPA (Roemmele, Bejan, & Gordon, 2011)

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

a. **My neighbor invited me in.**

b. My neighbor left his house.

(B) JOCI (Zhang, Rudinger, Duh, & Van Durme, 2017)

Context: John was excited to go to the fair
Hypothesis: The fair opens.
Label: **5 (very likely)**

Context: Today my water heater broke
Hypothesis: A person looks for a heater.
Label: **4 (likely)**

Context: John's goal was to learn how to draw well
Hypothesis: A person accomplishes the goal.
Label: **3 (plausible)**

Context: Kelly was playing a soccer match for her University
Hypothesis: The University is dismantled.
Label: **2 (technically possible)**

Context: A brown-haired lady dressed all in blue denim sits in a group of pigeons.
Hypothesis: People are made of the denim.
Label: **1 (impossible)**

(C) ROCStories (Mostafazadeh et al., 2016)

Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee.

Ending:

a. **Tom asked Sheryl to marry him.**

b. He wiped mud off of his boot.

(D) AlphaNLI (Bhagavatula, Bras, Malaviya, Sakaguchi, Holtzman, Rashkin, Downey, Yih, & Choi, 2019)

Observation 1: There was ten feet of snow outside.
Observation 2: In all that time I was unable to check my mail.
Hypotheses:

a. **I couldn't open my door against a drift for 3 days.**

b. It took 10 minutes for the snow plow to come through.

(E) SWAG (Zellers, Bisk, Schwartz, & Choi, 2018)

He pours the raw egg batter into the pan. He

a. drops the tiny pan onto a plate

b. **lifts the pan and moves it around to shuffle the eggs.**

c. stirs the dough into a kite.

d. swirls the stir under the adhesive.

MULTIPLE TASKS

- Some benchmarks consist of several focused language processing or reasoning subtasks so that a diverse set of reading comprehension skills can be learned and tested in a consistent format. These benchmarks can be used as diagnostics to determine how a model performs in different areas.

bAbI

- The bAbI benchmark from Weston et al. (2016) consists of 20 prerequisite tasks, each with 1,000 examples for training and 1,000 for testing. Each task presents systems with a passage, then asks a reading comprehension question. Each task also focuses on a different type of reasoning or language processing task, allowing systems to learn basic skills one at a time. Tasks are as follows:

- 1. Single supporting fact
- 2. Two supporting facts
- 3. Three supporting facts
- 4. Two argument relations
- 5. Three argument relations
- 6. Yes/no questions
- 7. Counting
- 8. Lists/sets
- 9. Simple negation
- 10. Indefinite knowledge
- 11. Basic coreference
- 12. Conjunction
- 13. Compound coreference
- 14. Time reasoning
- 15. Basic deduction
- 16. Basic induction
- 17. Positional reasoning
- 18. Size reasoning
- 19. Path finding
- 20. Agent’s motivations

MULTIPLE TASKS

- The General Language Understanding Evaluation (GLUE) dataset consists of 9 language tasks, including single-sentence binary classification and 2- or 3-way entailment comparable to the dual tasks in RTE-4 and RTE-5. The GLUE tasks are recast or included directly from other benchmark data and corpora:
  1. Corpus of Linguistic Acceptability (CoLA)
  2. Stanford Sentiment Treebank (SST-2)
  3. Microsoft Research Paraphrase Corpus (MRPC)
  4. Quora Question Pairs (QQP)
  5. Semantic Textual Similarity Benchmark (STS-B)
  6. Multi-Genre Natural Language Inference (MNLI)
  7. Question Natural Language Inference (QNLI)
  8. Recognizing Textual Entailment (RTE), RTE-2,RTE-3, and RTE-5
  9. Winograd Natural Language Inference (WNLI)

Task	bAbI	IIE	GLUE	DNC	SuperGLUE
Semantic Role Labeling		✓			
Relation Extraction	✓			✓	
Event Factuality				✓	
Named Entity Recognition				✓	
Word Sense Disambiguation					✓
Reference Resolution	✓	✓	✓	✓	✓
Grammaticality			✓		
Lexicosyntactic Inference				✓	✓
Sentiment Analysis			✓	✓	
Figurative Language				✓	
Sentence Similarity			✓		
Paraphrase		✓	✓	✓	
Sentence Completion					✓
Textual Entailment			✓	✓	✓
Question Answering	✓		✓		✓

TASK FORMAT

- Classification tasks
- Open-ended tasks

Each task is recast from preexisting data into classic RTE format

(A) GLUE, Question Answering NLI

Context: Who was the main performer at this year's halftime show?

Hypothesis: The Super Bowl 50 halftime show was headlined by the British rock group Coldplay with special guest performers Beyoncé and Bruno Mars, who headlined the Super Bowl XLVII and Super Bowl XLVIII halftime shows, respectively.

Label: **entailed**

(B) IIE, Definite Pronoun Resolution

Context: The bird ate the pie and it died.

Hypothesis: The bird ate the pie and the bird died.

Label: **entailed**

(C) DNC, Figurative Language

Context: Carter heard that a gardener who moved back to his home town rediscovered his roots.

Hypothesis: Carter heard a pun.

Label: **entailed**

(A) CBT (Hill et al., 2015)

1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . "
7 " Are the boys big ? "
8 queried Esther anxiously .
9 " Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em – that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . "
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best .
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .
21 She thought that Mr. _____ had exaggerated matters a little .

Blank: **Baxter**, Cropper, Esther, course, fingers, manner, objection, opinions, right, spite

(B) ROCStories (Mostafazadeh et al., 2016)

Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.

Ending:

a. **Karen became good friends with her roommate.**

b. Karen hated her roommate.

(C) CLOTH (Xie et al., 2017)

She pushed the door open and found nobody there. "I am the _____ to arrive." She thought and came to her desk.

a. last

b. second

c. third

d. **first**

(D) SWAG (Zellers et al., 2018)

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.**

(E) ReCoRD (Zhang et al., 2018)

... Daniela Hantuchova knocks Venus Williams out of Eastbourne 6-2 5-7 6-2 ...

Query: Hantuchova breezed through the first set in just under 40 minutes after breaking Williams' serve twice to take it 6-2 and led the second 4-2 before _____ hit her stride.

Venus Williams

(A) SQuAD 2.0 (Rajpurkar et al., 2018)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

(B) bAbI (Weston et al., 2016)

The kitchen is north of the hallway.
The bathroom is west of the bedroom.
The den is east of the hallway.
The office is south of the bedroom.

How do you go from den to kitchen?
west, north

(C) SC (Rashkin et al., 2018a)[9]

Valerie was getting ready for a formal dance. She had been preparing for hours. As she was ready to leave, her acrylic nail broke. She snapped off all of her faux nails.

Maslow: **esteem, stability**
Reiss: **status, approval, order**
Plutchik: **surprise, sadness, disgust, anger**

(D) DROP (Dua et al., 2019)

That year, his Untitled (1981), a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate.

Question: How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?
4300000

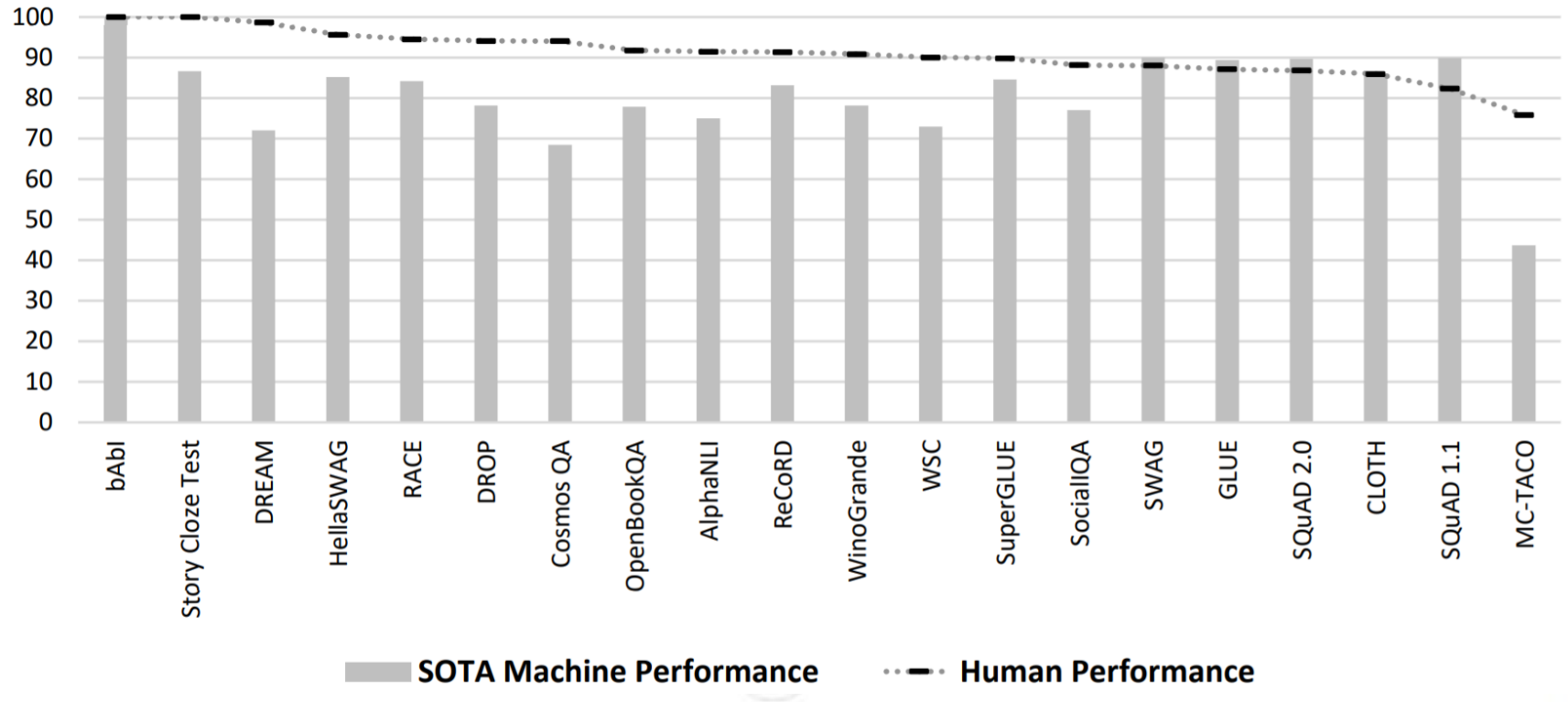
DATA BIASES

When creating benchmarks, one challenge is the bias of data unintentionally introduced to the benchmark. For example, in the first release of the Visual Question Answering (VQA) benchmark (Agrawal, Lu, Antol, Mitchell, Zitnick, Parikh, & Batra, 2017), researchers found that machine learning models were learning several statistical biases in the data, and could answer up to 48% of questions in the validation set without seeing the image.

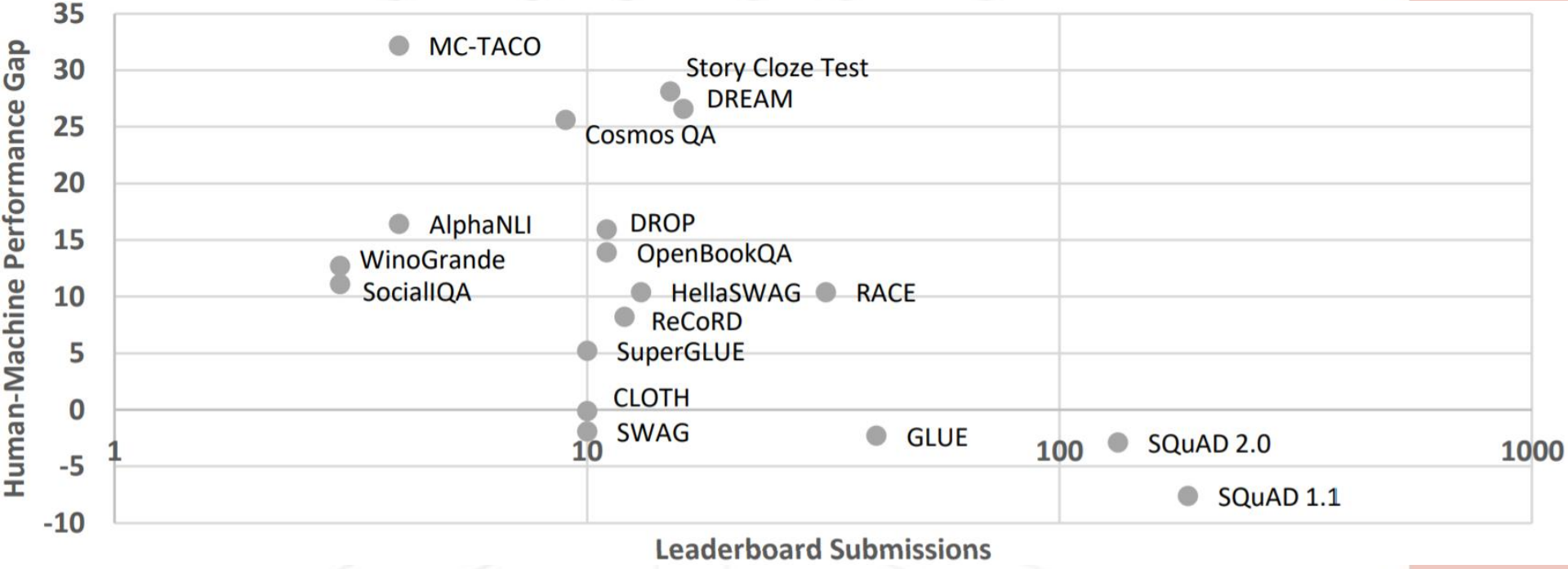
- Label distribution bias
- Question type bias
- Superficial correlation bias
- Addressing superficial correlation bias

EVALUATION SCHEMES

- Evaluation metrics
- Baseline performance
- Human performance measurement



Number of leaderboard submissions versus gap between state-of-the-art machine accuracy and human accuracy for selected benchmarks where human performance is reported and public leaderboards are maintained.



Knowledge Resources

- To understand human language, it is important to have linguistic knowledge resources that allow computers to identify syntactic and semantic structures from language. These structures often need to be augmented with common knowledge and commonsense knowledge to reach a full understanding.

- The acquired knowledge is often represented in various forms such as propositions, taxonomies, ontologies, and semantic networks:
  - LINGUISTIC KNOWLEDGE RESOURCES
  - COMMON KNOWLEDGE RESOURCES
  - COMMONSENSE KNOWLEDGE RESOURCES

Knowledge Resources

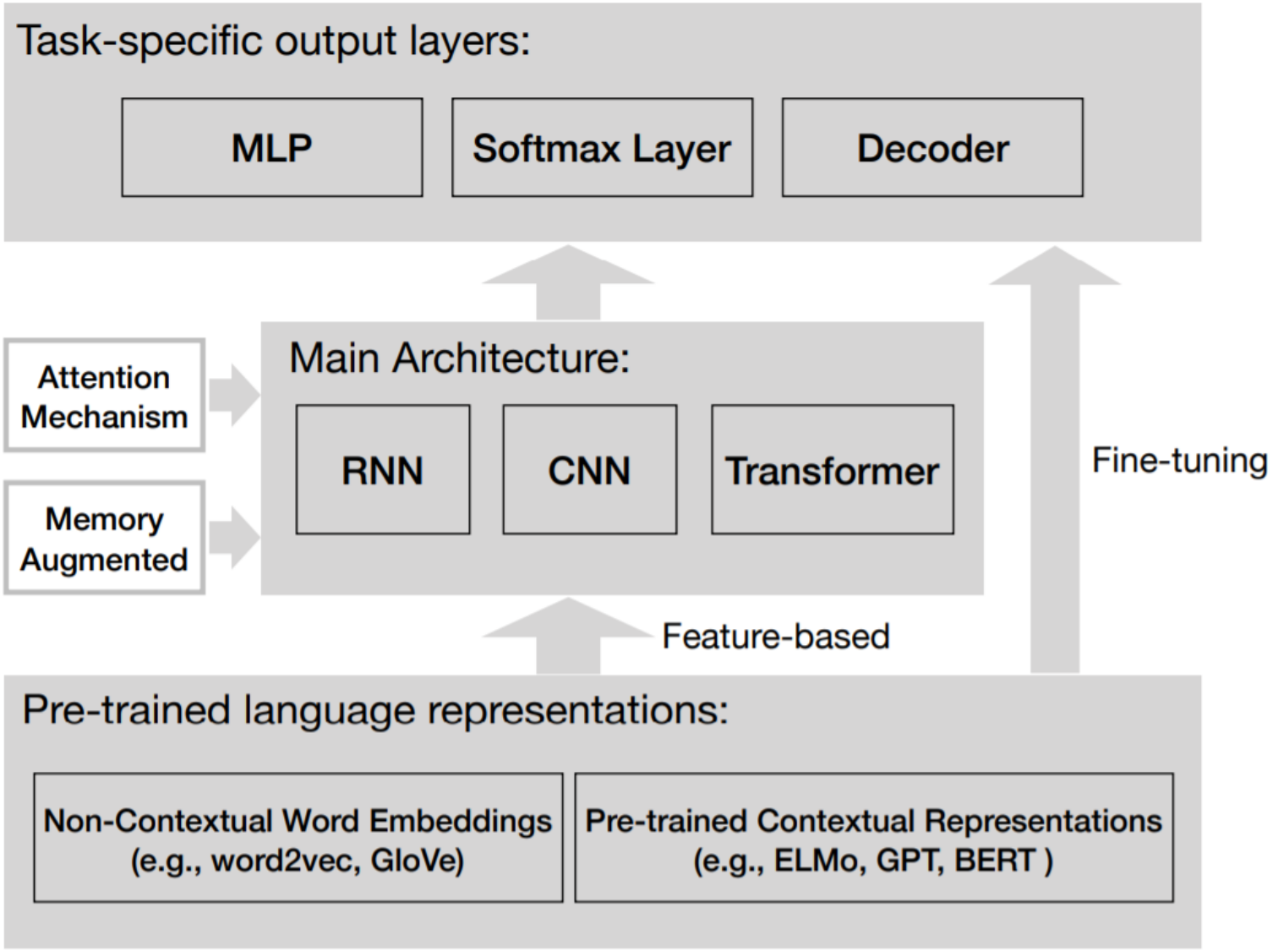
- LINGUISTIC KNOWLEDGE RESOURCES
  - Annotated linguistic corpora
  - Lexical resources.
  - Frame semantics.
  - Pre-trained semantic vectors
- COMMON KNOWLEDGE RESOURCES
  - YAGO
  - Dbpedia
  - Freebase
  - NELL
  - Wikidata
- COMMONSENSE KNOWLEDGE RESOURCES
  - Cyc
  - ConceptNet
  - ATOMIC

Learning and Inference Approaches

- To solve the benchmark tasks described in Section 2, a variety of approaches have been developed. These range from earlier symbolic and statistical approaches to recent approaches that apply deep learning and neural networks.
  - Symbolic Approaches
  - Statistical Approaches
  - Neural Approaches

Neural Approaches

- ATTENTION MECHANISM
- MEMORY AUGMENTATION
- CONTEXTUAL MODELS AND REPRESENTATIONS


Task-specific output layers:
MLP
Softmax Layer
Decoder
Attention Mechanism
Memory Augmented
Main Architecture:
RNN
CNN
Transformer
Fine-tuning
Feature-based
Pre-trained language representations:
Non-Contextual Word Embeddings (e.g., word2vec, GloVe)
Pre-trained Contextual Representations (e.g., ELMo, GPT, BERT)

Comparison of exact-match accuracy achieved on selected benchmarks by a random or majority-choice baseline, various neural contextual embedding models, and humans. ELMO refers to the highest-performing listed approach using ELMO embeddings. Best system performance on each benchmark in bold.

Benchmark	Simple Baseline	ELMo	GPT	BERT	MT-DNN	XLNET	ROBERTA	ALBERT	Human
CLOTH	25.0	70.7	–	**86.0**	–	–	–	–	85.9
Cosmos QA	–	–	54.5	67.1	–	–	–	–	94.0
DREAM	33.4	59.5	55.5	66.8	–	**72.0**	–	–	95.5
GLUE	–	70.0	–	80.5	87.6	88.4	88.5	**89.4**	87.1
HellaSWAG	25.0	33.3	41.7	47.3	–	–	**85.2**		95.6
MC-TACO	17.4	26.4	–	42.7	–	–	**43.6**	–	75.8
RACE	24.9	–	59.0	72.0	–	81.8	83.2	**89.4**	94.5
SciTail	60.3	–	88.3	–	94.1	–	–	–	–
SQuAD 1.1	1.3	81.0	–	87.4	–	**89.9**	–	–	82.3
SQuAD 2.0	48.9	63.4	–	80.8	–	86.3	86.8	**89.7**	86.9
SuperGLUE	47.1	–	–	69.0	–	–	**84.6**	–	89.8
SWAG	25.0	59.1	78.0	86.3	87.1	–	**89.9**	–	88.0

Incorporating External Knowledge

- Use of linguistic resources
  - WordNet
- Use of common knowledge resources
  - Dbpedia
  - YAGO
- Use of commonsense knowledge resources
  - ConceptNet
  - Cyc

Conclusion

- Need greater emphasis on external knowledge acquisition and incorporation
- Need greater emphasis on reasoning
- Need stronger justification and better understanding on design choices of models
- Need broader and multidimensional metrics for evaluation

Thank You