



Common NLP Tasks and Metrics

CS 780/880 Natural Language Processing Lecture 10

Samuel Carton, University of New Hampshire

Last lecture

Key idea: Hidden Markov Models

Concepts:

- Bayes Networks
- Generative story
- HMMs
 - Inference
 - Likelihood: **Forward algorithm**
 - Decoding: **Viterbi algorithm**
 - Learning
 - Labeled: Counting
 - Unlabeled: **Forward-backward algorithm**
 - Generation

Concepts:

- POS tagging
- Dynamic programming
- Expectation-maximization



Tools in your toolkit

Classification: given some labeled texts, create a model that can predict labels for new texts

- K-means clustering, Naïve Bayes

Word similarity: given two pairs of texts, assess which pair is more similar

- Word count vectors, TF-IDF vectors

Language modeling: given a corpus of texts, learn how to generate new texts and assess the likelihood of existing texts under the model

- Unigram model, bigram model

Sequence tagging (sort of): Given a corpus of texts and labels for each word, learn how to predict word labels for new texts

- Hidden Markov Models, ...?



GPT-3

Really just a big language model

Optimized to predict $p(w_i | w_{i-1}, w_{i-2}, w_{i-3}, \dots, w_0)$

- Remember that the best we've been able to accommodate is $p(w_i | w_{i-1})$ with our bigram model

Published in: <https://arxiv.org/abs/2005.14165>

Can do the three things we do with language models:

- Learning (already done)
- Inference (i.e assessing the likelihood of existing text)
- Generation



NLP dataset/task ecosystem

There are a lot of different tasks that NLP models are designed to do

- https://en.wikipedia.org/wiki/Natural_language_processing#Common_NLP_tasks

But these tasks are defined specifically by **datasets** put together by companies and researchers

The quality of NLP models is often defined by their ability to perform well on **benchmark collections** of these datasets



3.1) Language modeling, Cloze and Completion



Language modeling

Basic idea: Take a well-known corpus (other than what your model was trained on) and calculate **perplexity** of your language model that corpus

- A better model is one more likely to have generated that corpus (lower perplexity)

Metric: Perplexity—average per-word log-likelihood of words in the corpus, per the model

- N: corpus size, d_i : document i

$$\frac{1}{N} \sum_i^N \frac{\log(p(d_i))}{|d_i|}$$

Dataset: Penn TreeBank (PTB)

- 1 million words of 1989 Wall Street Journal article text
- <https://catalog ldc.upenn.edu/LDC99T42>

Question: Why might perplexity not be an ideal metric for what constitutes a good language model?

GPT-3 result: Improved SOTA by 1.5 down to 20.50



Completion/clozure

Basic idea: Test the language model with “gaps” for it to fill in, where getting the right answer involves “understanding” the text

Metric: Classification metrics (accuracy, P/R/F1)

- Count completions it did correctly versus not

Alice was friends with Bob. Alice went to visit her friend _____. → Bob
George bought some baseball equipment, a ball, a glove, and a _____. →

Dataset: LAMBADA

- https://zenodo.org/record/2630551#.Y_esCnZKhhE
- ~10,000 short passages, goal is to predict the last word

GPT-3 result: 8% improvement over previous SOTA

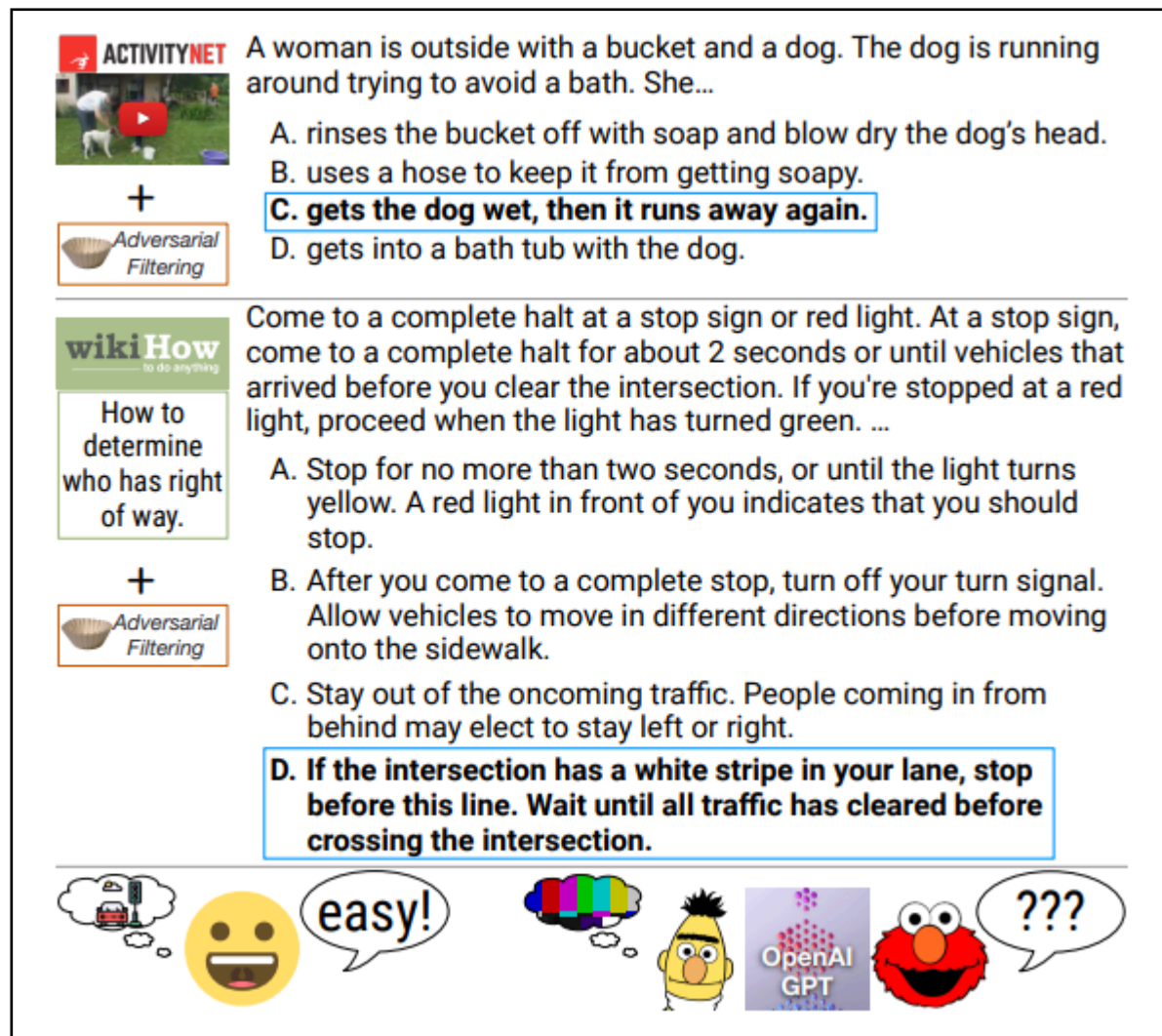


Completion/clozure

Dataset: HellaSwag

- <https://allenai.org/data/hellaswag>
- 70k instances
- Beginning of a story, plus multiple options for

Question: This isn't strictly a generation task like LAMBADA... so how do we apply a LM to this?



ACTIVITYNET A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...

A. rinses the bucket off with soap and blow dry the dog's head.
B. uses a hose to keep it from getting soapy.
C. gets the dog wet, then it runs away again.
D. gets into a bath tub with the dog.

wikiHow Come to a complete halt at a stop sign or red light. At a stop sign, come to a complete halt for about 2 seconds or until vehicles that arrived before you clear the intersection. If you're stopped at a red light, proceed when the light has turned green. ...

A. Stop for no more than two seconds, or until the light turns yellow. A red light in front of you indicates that you should stop.
B. After you come to a complete stop, turn off your turn signal. Allow vehicles to move in different directions before moving onto the sidewalk.
C. Stay out of the oncoming traffic. People coming in from behind may elect to stay left or right.
D. If the intersection has a white stripe in your lane, stop before this line. Wait until all traffic has cleared before crossing the intersection.

easy!

OpenAI GPT

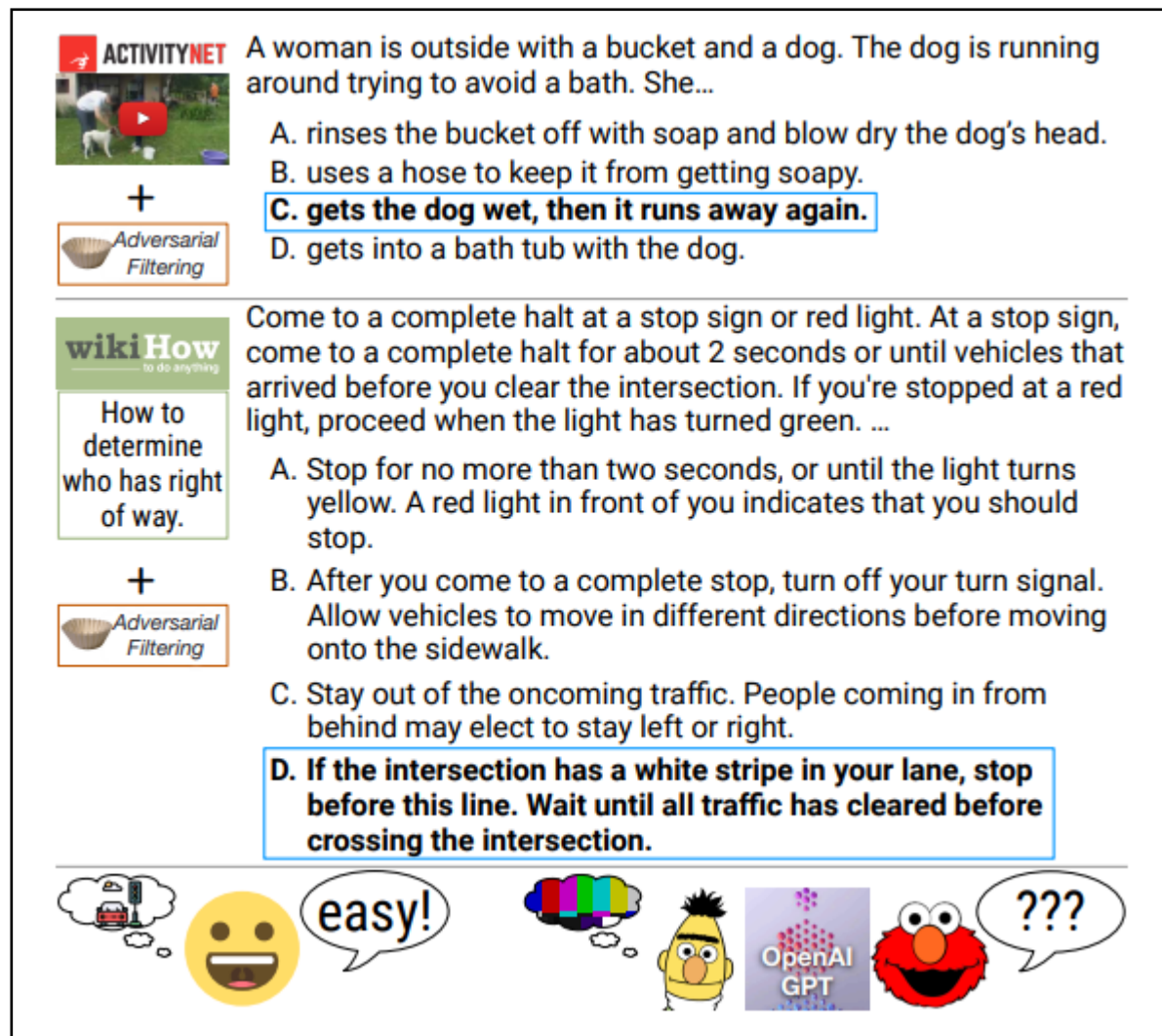
???

Completion/clozure

Answer: See which possible answer is most likely under the language model, and select that as the most likely one.

Metric: Classification metrics

GPT-3 result: less than SOTA



ACTIVITYNET A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...

- A. rinses the bucket off with soap and blow dry the dog's head.
- B. uses a hose to keep it from getting soapy.
- C. gets the dog wet, then it runs away again.**
- D. gets into a bath tub with the dog.

wikiHow Come to a complete halt at a stop sign or red light. At a stop sign, come to a complete halt for about 2 seconds or until vehicles that arrived before you clear the intersection. If you're stopped at a red light, proceed when the light has turned green. ...

- A. Stop for no more than two seconds, or until the light turns yellow. A red light in front of you indicates that you should stop.
- B. After you come to a complete stop, turn off your turn signal. Allow vehicles to move in different directions before moving onto the sidewalk.
- C. Stay out of the oncoming traffic. People coming in from behind may elect to stay left or right.
- D. If the intersection has a white stripe in your lane, stop before this line. Wait until all traffic has cleared before crossing the intersection.**

Thought bubble with car icon, Happy emoji, "easy!", Rainbow thought bubble, Yellow character, OpenAI GPT logo, Elmo, "???"

Completion/Clozure

Dataset: StoryCloze

- <https://cs.rochester.edu/nlp/rocstories/>
- 3,744 short stories with a “right ending” and a “wrong ending”

Context	Right Ending	Wrong Ending
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.	Tom asked Sheryl to marry him.	He wiped mud off of his boot.
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.	Karen became good friends with her roommate.	Karen hated her roommate.
Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a \$10,000 debt. Jim realized that he was foolish to spend so much money.	Jim decided to devise a plan for repayment.	Jim decided to open another credit card.

Table 4: Three example Story Cloze Test cases, completed by our crowd workers.



Closed book question answering

Basic idea: Literally just ask the LM a question and see if it generates the right answer

- As opposed to “open-book”, where it retrieves & uses external information (more on this later)

Metric: Classification metrics

Dataset: TriviaQA

- <https://nlp.cs.washington.edu/triviaqa/>
- 650k question-answer-evidence triplets

GPT-3 result: close to fine-tuned SOTA

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?

Answer: The Guns of Navarone

Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

Question: American Callan Pinckney’s eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?

Answer: Fitness

Excerpt: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became international best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney’s first video release “Callanetics: 10 Years Younger In 10 Hours” outsold every other **fitness** video in the US.



3.3) Translation



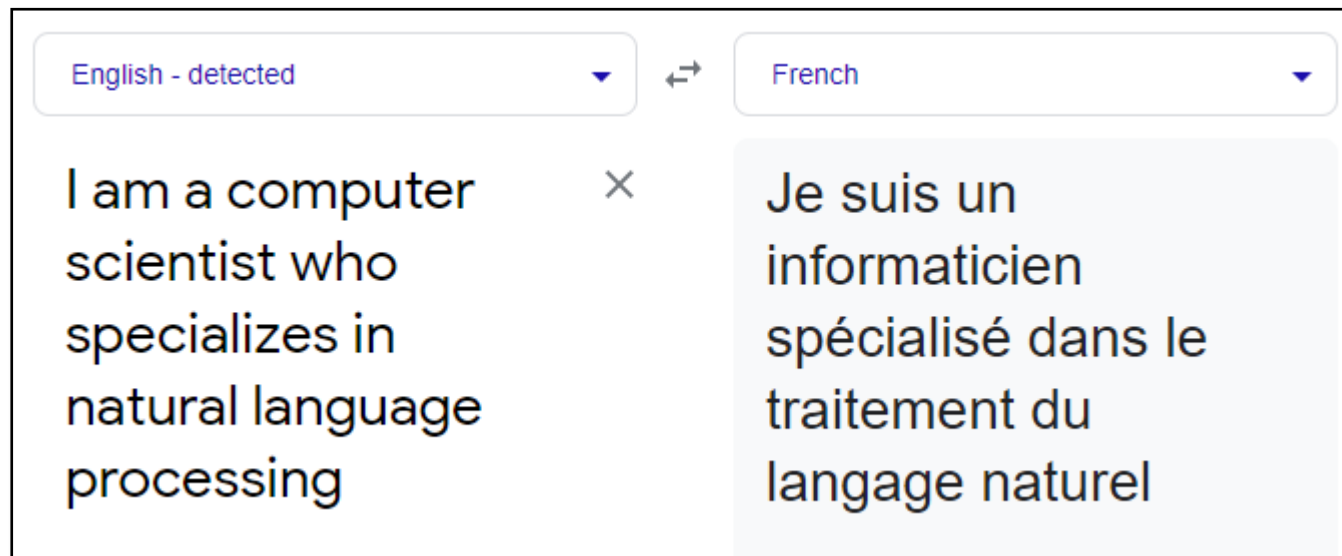
Machine translation

Basic idea: Given a text in one language, generate a text in a different language that has the same meaning

Example of a text-to-text (aka sequence-to-sequence) learning task, which you **have not learned how to do yet**.

Represents a whole genre of NLP research.

Difficult to do, difficult to evaluate



Google translate



Translation correspondences

Lots of different ways that translation can transform source text to target text.

We will learn how to do MT a little later.

For now, let's focus on **evaluation**

One to-one:

John loves Mary.
| | |
Jean aime Marie.

**One-to-many:
(and reordering)**

John told Mary a story.
| / | | |
Jean [a raconté] une histoire [à Marie]

**Many-to-one:
(and elision)**

John is a [computer scientist].
| | / |
Jean est informaticien.

Many-to-many:

John [swam across] the lake.
| / | | |
Jean [a traversé] le lac [à la nage].



Evaluating machine translation

Difficult because there can be **multiple valid target translations** of the same input text

Booz endormi (Original French - 1859-83 Victor Hugo)	Boaz Asleep (Translation circa late 1800s, various publishers)	Boaz Asleep (Translation - 2001 EH and AM Blackmore)	Boaz Asleep (translation - 2002 Brooks Haxton)
<p>Booz s'était couché de fatigue accablé; Il avait tout le jour travaillé dans son aire; Puis avait fait son lit à sa place ordinaire; Booz dormait auprès des boisseaux pleins de blé. Ce vieillard possédait des champs de blés et d'orge; Il était, quoique riche, à la justice enclin; Il n'avait pas de fange en l'eau de son moulin; Il n'avait pas d'enfer dans le feu de sa forge.</p>	<p>At work within his barn since very early, Fairly tired out with toiling all the day, Upon the small bed where he always lay Boaz was sleeping by his sacks of barley. Barley and wheat fields he possessed, and well, Though rich, loved justice; wherfore all the flood That turned his mill-wheels was unstained with mud, And in his smithy blazed no fire of hell.</p>	<p>There Boaz lay, overcome and worn out. All day he'd labored at his threshing floor; Now, bedded in his usual place once more, He slept, with grain bagged everywhere about. Boaz owned fields of barleycorn and wheat-- A rich old man, but righteous, even so. There was no foulness in his millstream's flow, There was no hellfire in his forge's heat.</p>	<p>Boaz, overcome with weariness, by torchlight made his pallet on the thresing floor where all day he had worked, and now he slept among the bushels of threshed wheat. The old man owned wheatfields and barley, and though he was rich, he was still fair-minded. No filth soured the sweetness of his well. No hot iron of torture whitened his forge.</p>

<http://www.gavroche.org/vhugo/vhpoetry/comparison.gav>



BLEU score

Basic idea: Provide several reference target texts, and measure how well the model matched any/all of them

- Not idea, but works pretty well in practice

Based on **N-gram precision**: how many n-grams in the candidate translation occur also in one of the reference translations?

C1: It is a guide to action which ensures that the military always obeys the commands of the party.

C2: It is to insure the troops forever hearing the activity guidebook that party direct

R1: It is a guide to action that ensures that the military will forever heed Party commands.

R2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

R3: It is the practical guide for the army always to heed the directions of the party.

BLEU details

For $n \in \{1, \dots, 4\}$, compute the (modified) **precision of all n -grams**:

$$Prec_n = \frac{\sum_{c \in C} \sum_{n\text{-gram} \in c} \text{MaxFreq}_{\text{ref}}(n\text{-gram})}{\sum_{c \in C} \sum_{n\text{-gram} \in c} \text{Freq}_c(n\text{-gram})}$$

$\text{MaxFreq}_{\text{ref}}('the party')$ = max. count of *'the party'* in **one** reference translation.

$\text{Freq}_c('the party')$ = count of *'the party'* in candidate translation c .

Penalize short candidate translations by a **brevity penalty BP**

c = length (number of words) of the whole candidate translation corpus

r = Pick for each candidate the reference translation that is closest in length;
sum up these lengths.

Brevity penalty $BP = \exp(1 - c/r)$ for $c \leq r$; $BP = 1$ for $c > r$

(BP ranges from e for $c=0$ to 1 for $c=r$)



GPT-3 translation results

Several different translation datasets.

Worse than SOTA, but pretty good considering it's not actually trained to do translation

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro	Ro→En
SOTA (Supervised)	45.6^a	35.0 ^b	41.2^c	40.2 ^d	38.5^e	39.9^e
XLM [LC19]	33.4	33.3	26.4	34.3	33.3	31.8
MASS [STQ ⁺ 19]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>	33.1
mBART [LGG ⁺ 20]	-	-	<u>29.8</u>	34.0	<u>35.0</u>	30.5
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1	19.9
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6	38.6
GPT-3 Few-Shot	32.6	<u>39.2</u>	29.7	<u>40.6</u>	21.0	<u>39.5</u>



3.4) Winograd-style tasks



Winograd Schema Challenge

Basic idea: Create a sentence with a pronoun that is ambiguous based on the choice of the following verb, then see if the model is able to correctly disambiguate it when each choice is selected.

Metric: Classification metrics (acc, P/R/F1)

Question: How can we get a language model to do this?

The first cited example of a Winograd schema (and the reason for their name) is due to [Terry Winograd](#):^[7]

The city councilmen refused the demonstrators a permit because **they** [feared/advocated] violence.

The choices of "feared" and "advocated" turn the schema into its two instances:

The city councilmen refused the demonstrators a permit because **they** feared violence.

The city councilmen refused the demonstrators a permit because **they** advocated violence.

Winograd Schema Challenge

Answer: See which complete disambiguation is more likely under the model

Example: The city councilmen refused the demonstrators a permit because **the city councilmen** advocated violence
vs.
The city councilmen refused the demonstrators a permit because **the demonstrators** advocated violence

GPT-3 result: Close to SOTA

The first cited example of a Winograd schema (and the reason for their name) is due to [Terry Winograd](#):^[7]

The city councilmen refused the demonstrators a permit because **they** [feared/advocated] violence.

The choices of "feared" and "advocated" turn the schema into its two instances:

The city councilmen refused the demonstrators a permit because **they** feared violence.

The city councilmen refused the demonstrators a permit because **they** advocated violence.

3.5) Common- sense reasoning



Common-sense reasoning

Basic idea: comprehension tasks which **also** rely on external knowledge

Dataset: PhysicalQA (PIQA)

- Common sense questions with a right answer and a wrong answer
- ~20k examples

Metric: Classification metrics (acc, P/R/F1)

GPT-3 result: Improves on fine-tuned SOTA (!)

The diagram is enclosed in a black rectangular border. At the top left is a small icon of a mountain. To its right is a light blue rectangular box containing the text: "To separate egg whites from the yolk using a water bottle, you should...". Below this are two more light blue boxes, labeled 'a.' and 'b.'. Box 'a.' contains the text: "a. **Squeeze** the water bottle and press it against the yolk. **Release**, which creates suction and lifts the yolk." Box 'b.' contains the text: "b. **Place** the water bottle and press it against the yolk. **Keep pushing**, which creates suction and lifts the yolk." Below box 'a.' is a yellow smiley face emoji and a small blue speech bubble containing the letter 'a!'. Above the smiley face is a thought bubble containing a photograph of a hand using a water bottle to separate an egg yolk. Below box 'b.' is a blue robot icon with a speech bubble containing three question marks '???'.

<https://arxiv.org/pdf/1911.11641.pdf>



Common-sense reasoning

Dataset: OpenBookQA

- <https://allenai.org/data/open-book-qa>
- ~6k multiple-choice science questions

Metric: Classification metrics (acc, P/R/F1)

GPT-3 result: Significantly worse than SOTA

Question:

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.

Science Fact:

Metal is a thermal conductor.

Common Knowledge:

Steel is made of metal.

Heat travels through a thermal conductor.

Figure 1: An example for a question with a given set of choices and supporting facts.



3.6) Reading comprehension



Reading comprehension

Basic idea: Answer questions about a text that require understanding of the text

- CoQA: <https://stanfordnlp.github.io/coqa/>
- DROP: <https://allenai.org/data/drop>
- QUAC: <https://quac.ai/>
- SQuADv2: <https://rajpurkar.github.io/SQuAD-explorer/>
- RACE: <https://www.cs.cmu.edu/~glai1/data/race/>

Metric: Classification (acc, P/R/F1)

GPT-3 results: mixed, but all well below fine-tuned SOTA

Question: What's the difference between this and the completion/clozure tasks from before?



Reading comprehension

Dataset: CoQA

- <https://stanfordnlp.github.io/coqa/>
- 127k questions with answers about 8k passages

Question: How to apply a language model to this dataset?

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q₁: Who had a birthday?

A₁: Jessica

R₁: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q₂: How old would she be?

A₂: 80

R₂: she was turning 80

Q₃: Did she plan to have any visitors?

A₃: Yes

R₃: Her granddaughter Annie was coming over

Reading comprehension

Dataset: SQuADv2

- <https://rajpurkar.github.io/SQuAD-explorer/>
- 50k difficult-to-answer questions about passages from original SQuAD dataset

Question: How to apply a language model to this dataset?

Article: Endangered Species Act

Paragraph: “ ... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.”

Question 1: “Which laws faced significant opposition?”

Plausible Answer: later laws

Question 2: “What was the name of the 1937 treaty?”

Plausible Answer: Bald Eagle Protection Act

Figure 1: Two unanswerable questions written by crowdworkers, along with plausible (but incorrect) answers. Relevant keywords are shown in blue.

3.7) SuperGLUE



SuperGLUE

<https://super.gluebenchmark.com/>

Basic idea: A collection of datasets designed to test the general language-understanding capability of an NLP model

Metric: multiple, but mostly just classification metrics

GPT-3 results: well below fine-tuned SOTA

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy

SuperGLUE

Dataset: MultiRC

- A series of paragraphs with reading comprehension questions and possible correct and correct answers
- <https://cogcomp.seas.upenn.edu/multirc/>
- ~6k questions about ~800 questions

Note: multiple possible correct answers

Question: how to apply a LM to this dataset?

S1: Most young mammals, including humans, play.
S2: Play is how they learn the skills that they will need as adults.
S6: Big cats also play.
S8: At the same time, they also practice their hunting skills.
S11: Human children learn by playing as well.
S12: For example, playing games and sports can help them learn to follow rules.
S13: They also learn to work together.

What do human children learn by playing games and sports?
A)* They learn to follow rules and work together
B) hunting skills
C)* skills that they will need as adult

Figure 1: Examples from our MultiRCcorpus. Each example shows relevant excerpts from a paragraph; multi-sentence question that can be answered by combining information from multiple sentences of the paragraph; and corresponding answer-options. The correct answer(s) is indicated by a *. Note that there can be multiple correct answers per question.

SuperGLUE

Dataset: BoolQ

- <https://github.com/google-research-datasets/boolean-questions>
- A series of yes-no reading comprehension questions based on short passages

Reasoning Types	Yes/No Question Answering Examples
Paraphrasing (38.7%) The passage explicitly asserts or refutes what is stated in the question.	Q: Is Tim Brown in the Hall of Fame? P: Brown has also played for the Tampa Bay Buccaneers. In 2015, he was inducted into the Pro Football Hall of Fame. A: Yes. [“inducted into” directly implies he is in Hall of Fame.]
By Example (11.8%) The passage provides an example or counter-example to what is asserted by the question.	Q: Are there any nuclear power plants in Michigan? P: ... three nuclear power plants supply Michigan with about 30% of its electricity. A: Yes. [Since there must be at least three.]
Factual Reasoning (8.5%) Answering the question requires using world-knowledge to connect what is stated in the passage to the question.	Q: Was designated survivor filmed in the White House? P: The series is... filmed in Toronto, Ontario. A: No. [The White House is not located in Toronto.]
Implicit (8.5%) The passage mentions or describes entities in the question in way that would not make sense if the answer was not yes/no.	Q: Is static pressure the same as atmospheric pressure? P: The aircraft designer’s objective is to ensure the pressure in the aircraft’s static pressure system is as close as possible to the atmospheric pressure... A: No. [It would not make sense to bring them “as close as possible” if those terms referred to the same thing.]



3.8) Natural Language Inference (NLI)



Natural language inference

Basic idea: Tests the model's ability to determine the relationship between two sentences

- E.g. does the first sentence **entail** the second sentence?

Dataset: Adversarial Natural Language Inference (ANLI)

- <https://github.com/facebookresearch/anli>
- ~200k context-hypothesis pairs where the context may or may not entail the hypothesis

Metric: Classification metrics (acc, P/R/F1)

Context	Hypothesis
Roberto Javier Mora García (c. 1962 – 16 March 2004) was a Mexican journalist and editorial director of “El Mañana”, a newspaper based in Nuevo Laredo, Tamaulipas, Mexico. He worked for a number of media outlets in Mexico, including the “El Norte” and “El Diario de Monterrey”, prior to his assassination.	Another individual laid waste to Roberto Javier Mora Garcia.
A melee weapon is any weapon used in direct hand-to-hand combat; by contrast with ranged weapons which act at a distance. The term “melee” originates in the 1640s from the French word “mêlée”, which refers to hand-to-hand combat, a close quarters battle, a brawl, a confused fight, etc. Melee weapons can be broadly divided into three categories	Melee weapons are good for ranged and hand-to-hand combat.

Concluding thoughts

Many:

- Tasks
- Datasets

Not so many:

- Basic ways of employing the model
- Evaluation metrics

NLP largely the art of adapting the ways we know how to use our models to specific linguistic tasks

GPT-3 is can do a lot with just LM training

