NLP Tasks, Data, and Evaluation

CS288

UC Berkeley

Today

- What tasks have NLP researchers *traditionally* cared about?

- How do we evaluate success?

  - Dominant paradigm: automatic metrics computed on static benchmark

- How do we collect benchmark datasets?

Modeling Linguistic Structure

- High-level question: how can language technologies represent and process language the way we think people do?

- Task: map from a piece of text to a linguistic formalism

- Evaluation is easy to perform automatically

- However, corpora often need to be carefully curated

Syntactic Parsing

Battle-tested/NNP*/JJ industrial/JJ managers/NNS here/RB always/RB buck/VB*/VBP up/IN*/RP nervous/JJ newcomers/NNS with/IN the/DT tale/NN of/IN the/DT first/JJ of/IN their/PP$ countrymen/NNS to/TO visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/NNS*/FW warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.
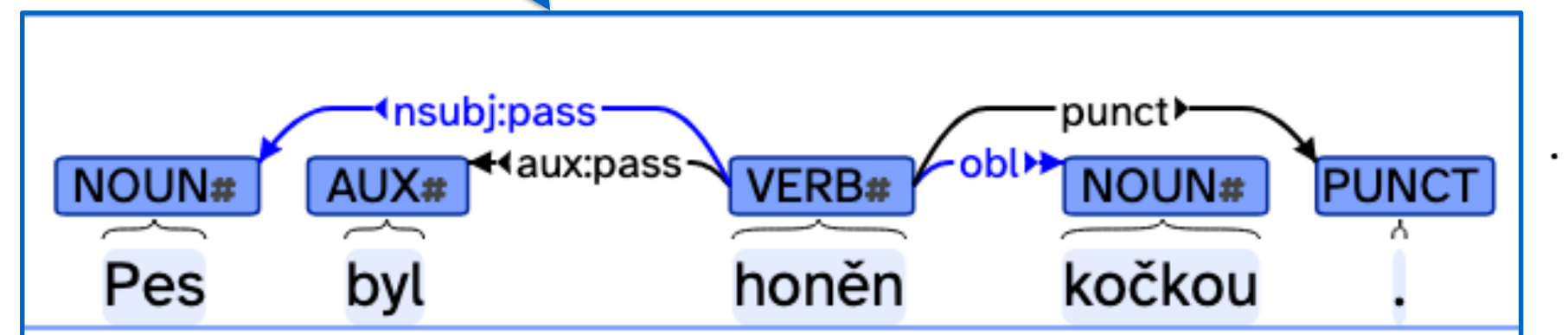"/" From/IN the/DT beginning/NN ,/, it/PRP took/VBD a/DT man/NN with/IN extraordinary/JJ qualities/NNS to/TO succeed/VB in/IN Mexico/NNP ,/, "/" says/VBZ Kimihide/NNP Takimura/NNP ,/, president/NN of/IN Mitsui/NNS*/NNP group/NN 's/POS Kensetsu/NNP Engineering/NNP Inc./NNP unit/NN ./.

POS tagging

```
# visual-style  3  1  nsubj:pass          color:blue
# visual-style  3  4  obl       color:blue
1        Pes        pes        NOUN        _    Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing
2        byl        být        AUX         _    Aspect=Imp|Gender=Masc|Number=Sing|Tense=Past|Ve
3        honěn      honit      VERB        _    Aspect=Imp|Gender=Masc|Number=Sing|VerbForm=Part
4        kočkou     kočka      NOUN        _    Case=Ins|Gender=Fem|Number=Sing 3
5        .          .          PUNCT       _    _                  3           punct
```
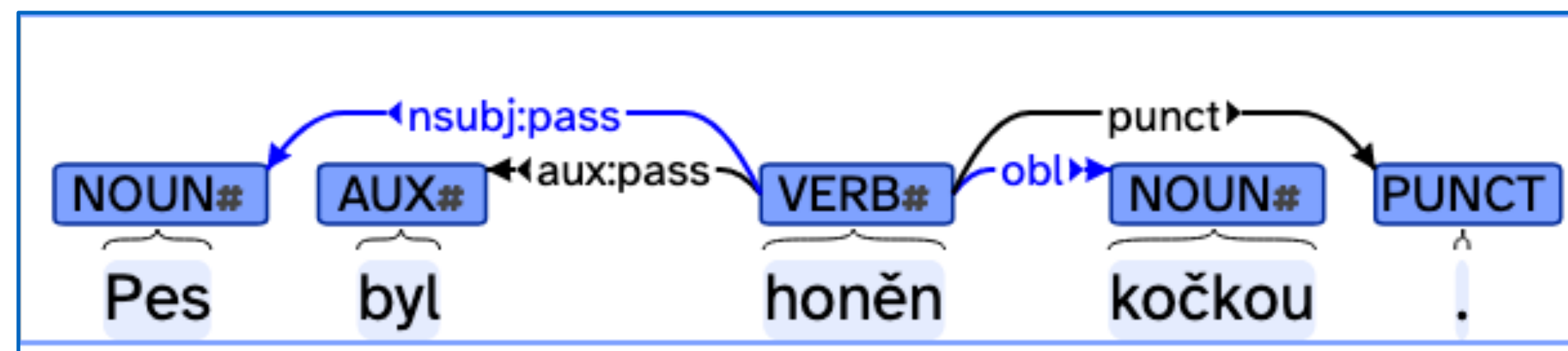
Dependency parsing

From Universal Dependencies,
https://universaldependencies.org/introduction.html

Constituency parsing

```
( (S
  (NP Battle-tested industrial managers
      here)
   always
  (VP buck
      up
      (NP nervous newcomers)
      (PP with
          (NP the tale
              (PP of
                  (NP (NP the
                      (ADJP first
                          (PP of
                              (NP their countrymen)))
                      (S (NP *)
                          to
                          (VP visit
                              (NP Mexico))))
                      ,
                  (NP (NP a boatload
                      (PP of
                          (NP (NP warriors)
                              (VP-1 blown
                                  ashore
                                  (ADVP (NP 375 years)
                                      ago)))))
                  (VP-1 *pseudo-attach*))))))))
```

.)

from the Penn Treebank,
Marcus et al. 1993

Syntactic Parsing: Evaluation Metrics

- Main intuition: how well did the model recover an underlying linguistic formalism for a particular sentence?

- Metrics:

  - Exact match (strict metric)

  - Attachment score (for dependency parsing) — what proportion of words are attached to the correct head?

  - Precision (accuracy among recovered constituents or tags), recall (what proportion of actual constituents or tags were recovered), and F1 (harmonic mean of these)

Syntactic Parsing: Corpora

- Need meticulously annotated corpora

- How to build a corpus?

  - Acquire source data

  - Develop an annotation scheme

  - Train annotators

  - Annotate the data! (May take a couple of years)

Acquiring Source Data

- Where do you get lots of text before the Internet was widely used?

Acquiring Source Data

- Where do you get lots of text before the Internet was widely used?

- Scanned documents: the Brown Corpus contains 1M words scanned from fiction and nonfiction prose in 1961

- Transcribed voice messages, e.g., radio broadcasts

- Multilingual data: often use the Bible; linguistics textbooks; transcripts of international meetings (e.g., EU)

- Early Internet: Wikipedia, blogs, reviews

Table 4
Penn Treebank (as of 11/92).

Marcus et al. 1992

Description	Tagged for Part-of-Speech (Tokens)	Skeletal Parsing (Tokens)
Dept. of Energy abstracts	231,404	231,404
Dow Jones Newswire stories	3,065,776	1,061,166
Dept. of Agriculture bulletins	78,555	78,555
Library of America texts	105,652	105,652
MUC-3 messages	111,828	111,828
IBM Manual sentences	89,121	89,121
WBUR radio transcripts	11,589	11,589
ATIS sentences	19,832	19,832
Brown Corpus, retagged	1,172,041	1,172,041
Total:	4,885,798	2,881,188

Icon in Figure 5	Genre
☁	bible
📅	blog
📔	fiction
✎	grammar examples
⚒	legal text
💉	medical text
🔲	news
ⓘ	non-fiction
👍	reviews
💬	spoken
📶	social (other user-generated content)
🌐	web
W	wikipedia

Nivre et al. 2016

Table 5: Genres present in the UD treebanks.

https://en.wikipedia.org/wiki/Brown_Corpus

Developing an Annotation Scheme

- Principles of developing a scheme
  - Simplicity
  - Consistent rules
  - Leaving room for ambiguity
  - Generalizability, to domains and across languages (this is hard!)
- This is very time-consuming
  - Build on top of existing schemes, identifying where they fail
  - Collaborate with others working with similar data to agree on standards
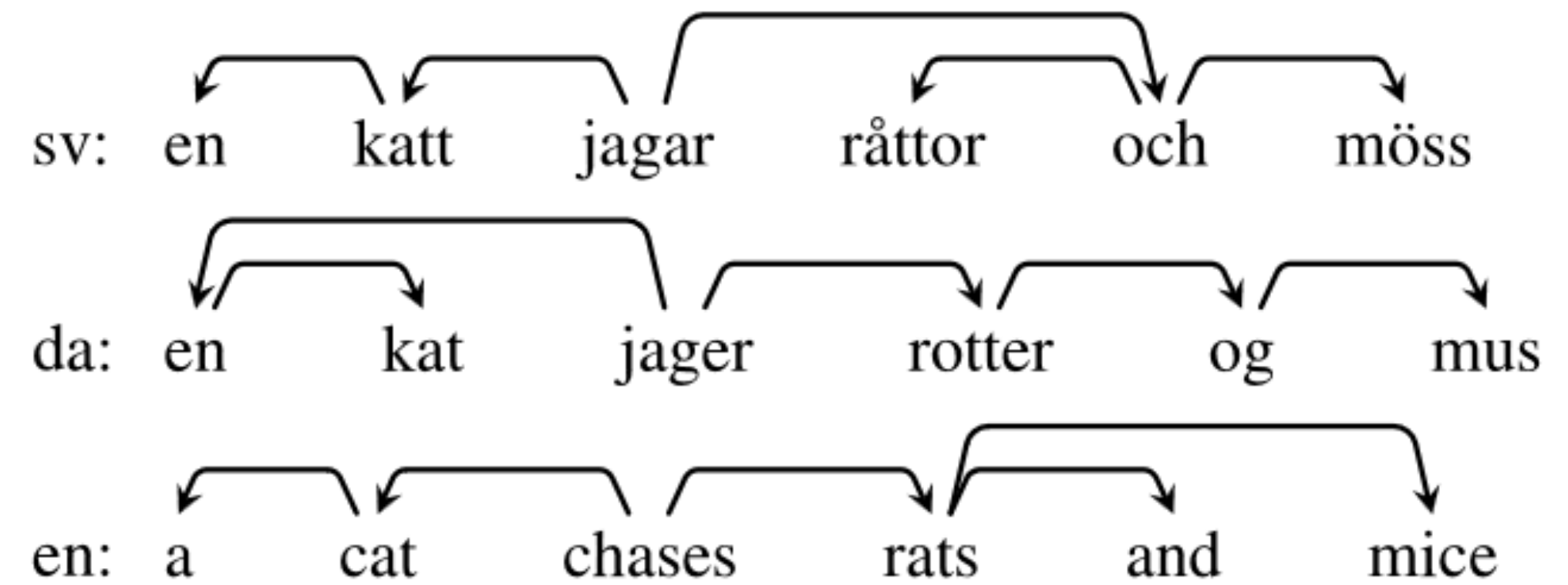- Also need a user-friendly interface for annotation

Figure 1: Divergent annotation of parallel structures

Dependency annotation
without standardization
(Nivre et al. 2016)

Annotation Process

- Training

  - Essentially need to train experts

  - Usually, they are people with at least some interest in the topic, e.g., linguists

  - Complex rules means annotation guidelines can be dozens of pages long!

- Managing annotators: need to quickly resolve disagreements and confusions

The annotators themselves were drawn from a variety of backgrounds, from undergraduates to holders of doctorates, including linguists, computer scientists, and others. Undergraduates have the advantage of being inexpensive but tend to work for only a few months each, so they require frequent training. Linguists make the best overall judgments although several of our nonlinguist annotators also had excellent skills. The learning curve for the annotation task tended to be very steep, with most annotators becoming comfortable with the process within three days of work. This

PropBank (Palmer et al. 2005) has a 66-page annotation guidelines document

Annotation Process

- Making it more efficient with automatic annotation

  - Use a smaller, less-performant model to perform the task

  - Then get human experts to identify and correct errors

- What about disagreement?

  - Several statistics measure inter-annotator agreement (e.g. via Cohen's kappa)

  - Low agreement could indicate that guidelines are confusing or inconsistent, the task is ambiguous or subjective, or carelessness

  - But low agreement should not be cast away as noise! Instead, we should carefully investigate sources of disagreement (see Leonardelli et al. 2021)

Aside: Using Corpora in Experiments

- Typically, with enough data, corpora are split into standard training / development / test sets:

    - Training: model should have full access

    - Development: should only be used for something like hyperparameter tuning or error analysis

    - Test:

        - Should be ran as infrequently as possible, to most closely model "real-world" model performance on unseen data

        - Sometimes not even publicly released!

Semantic Parsing

Broad-coverage semantic parsing

Abstract Meaning Representation, Banarescu et al. 2013

PropBank, Palmer et al. 2005

```
(d / describe-01
   :arg0 (m / man)
   :arg1 (m2 / mission)
   :arg2 (d / disaster))
```

	buy	sell
Arg0:	buyer	seller
Arg1:	thing bought	thing sold
Arg2:	seller	buyer
Arg3:	price paid	price paid
Arg4:	benefactive	benefactive

[Arg0 Chuck] *bought* [Arg1 a car] [Arg2 from Jerry] [Arg3 for $1000].

[Arg0 Jerry] *sold* [Arg1 a car] [Arg2 to Chuck] [Arg3 for $1000].

```
The man described the mission as a disaster.
The man's description of the mission:
   disaster.
As the man described it, the mission was a
   disaster.
```

Executable Semantic Parsing

GeoQuery, Tang and Mooney 2001

Input: "What is the largest city in Texas?"

Query: answer(C,largest(C,(city(C),loc(C,S),const(S,stateid(texas))))).

Spider, Yu et al. 2018

show me flights from seattle to boston next monday
```
(SELECT DISTINCT flight.flight_id FROM flight WHERE (flight.from_airport IN (SELECT
airport_service.airport_code FROM airport_service WHERE airport_service.city_code IN (SELECT
city.city_code FROM city WHERE city.city_name = 'SEATTLE'))) AND (flight.to_airport IN (SELECT
airport_service.airport_code FROM airport_service WHERE airport_service.city_code IN (SELECT
city.city_code FROM city WHERE city.city_name = 'BOSTON'))) AND (flight.flight_days IN (SELECT
days.days_code FROM days WHERE days.day_name IN (SELECT date_day.day_name FROM date_day WHERE
date_day.year = 1993 AND date_day.month_number = 2 AND date_day.day_number = 8))));
```

ATIS
Hemphill et al. 1990

What are the name and budget of the departments with average instructor salary greater than the overall average?

```
SELECT T2.name, T2.budget
FROM instructor as T1 JOIN department as
T2 ON T1.department_id = T2.id
GROUP BY T1.department_id
HAVING avg(T1.salary) >
   (SELECT avg(salary) FROM instructor)
```

13

Executable Semantic Parsing

- High-level goal: building a natural language interface that correctly answers user queries or executes their commands

- Often focuses on interfaces to databases

- Evaluation:

  - Can do exact match (but probably too strict!)

  - Evaluate denotational semantics with execution accuracy

Data Sources for Executable Semantic Parsing

- Ideally: should be questions and actions people actually would produce in-domain

- Frequently: utterances thought of on the fly, or even summaries of generated queries

SQL: select email_address from professionals where state = 'Hawaii' or state = 'Wisconsin';

Annotations

- Minimally: denotation (i.e., resulting answer).

  - This is much easier to annotate, but makes learning harder (requires weakly supervised learning, e.g., Krishnamurthy and Mitchell 2012)

- Ideally: also include logical form (e.g., SQL)

  - Doesn't (necessarily) require manual engineering of logical form, unlike syntactic formalisms

  - However, still requires expertise

Document Analysis and Understanding

- High level questions:

  - What information is encoded in a given document?

  - How can we combine information from a variety of documents?

- Encompasses a variety of tasks:

  - Document analysis, including topic modeling, document classification, etc.

  - Question answering — how well does a model represent a document such that you can ask reasonable questions about its contents?

  - Information extraction — can you summarize information across documents in a useful way?

Question Answering

- Given a document and a user question, produce a response

- Evaluation

  - Exact match of answers (maybe too strict)

  - N-gram overlap in generated answers (might still be too strict)

  - Multiple choice (might be too lenient)

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

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

From SQuAD, Rajpurkar et al. 2016

Data Collection and Annotation

- Data sources: usually nonfiction prose, e.g., from Wikipedia

- Crowdsourcing:

  - Workers are asked to come up with questions

  - Also collected additional answers from other workers to validate the labels

QA as an Evaluation Format

- Question answering is a widely used task format, with variations

- Multi-hop QA and reasoning-heavy QA (e.g., StrategyQA, Geva et al. 2021)

- Visual QA (Agrawal et al. 2015)

- Commonsense QA (e.g., Talmor et al. 2019)

*Where on a **river** can you hold a cup upright to catch water on a sunny day?*

✓ waterfall,　✗ bridge,　✗ valley,　✗ *pebble*,　✗ **mountain**

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
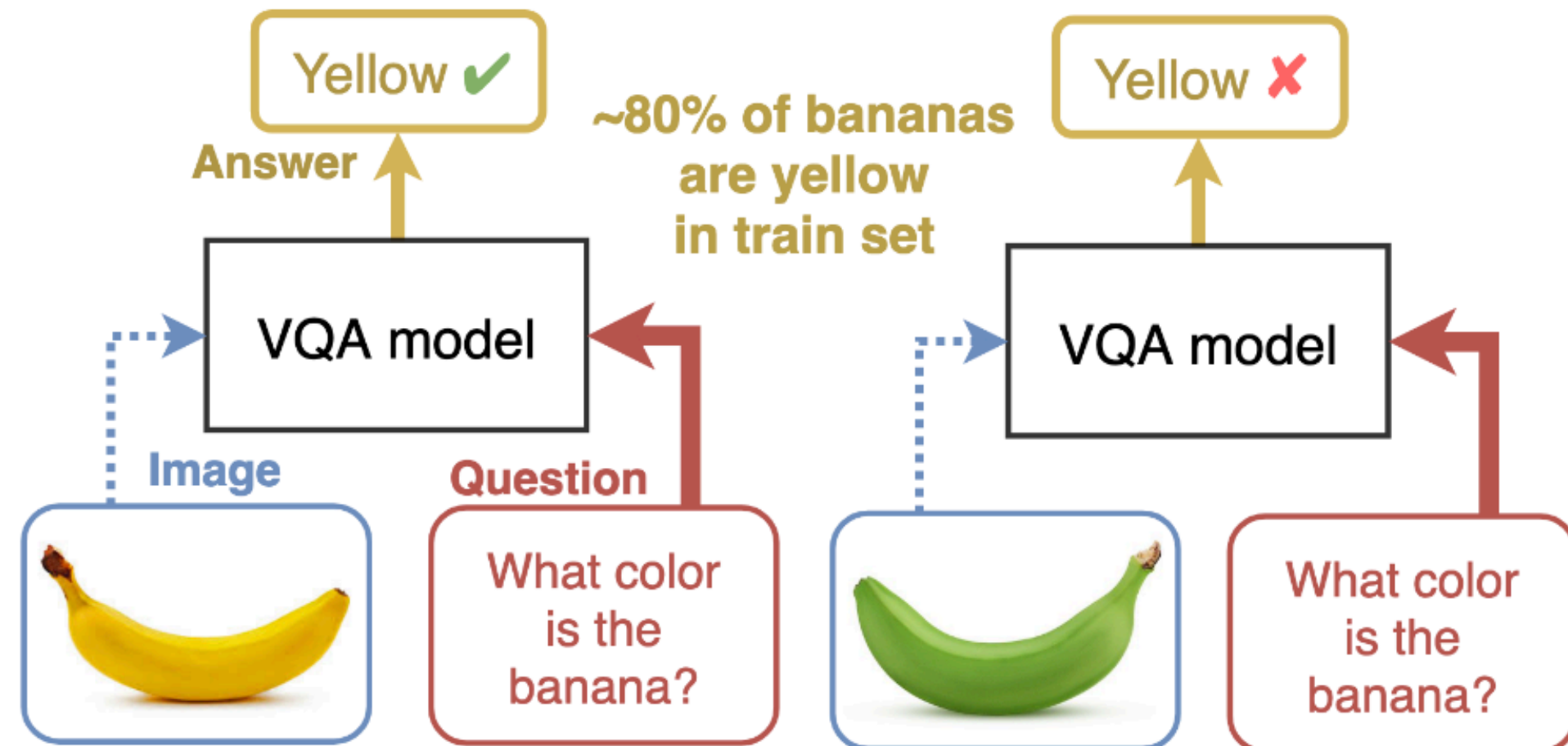**Supporting facts:** 1, 2, 4, 6, 7

From HotPotA, Yang et al. 2018

QA as an Evaluation Format

- But even with multiple choice, we still have to be careful about evaluation…

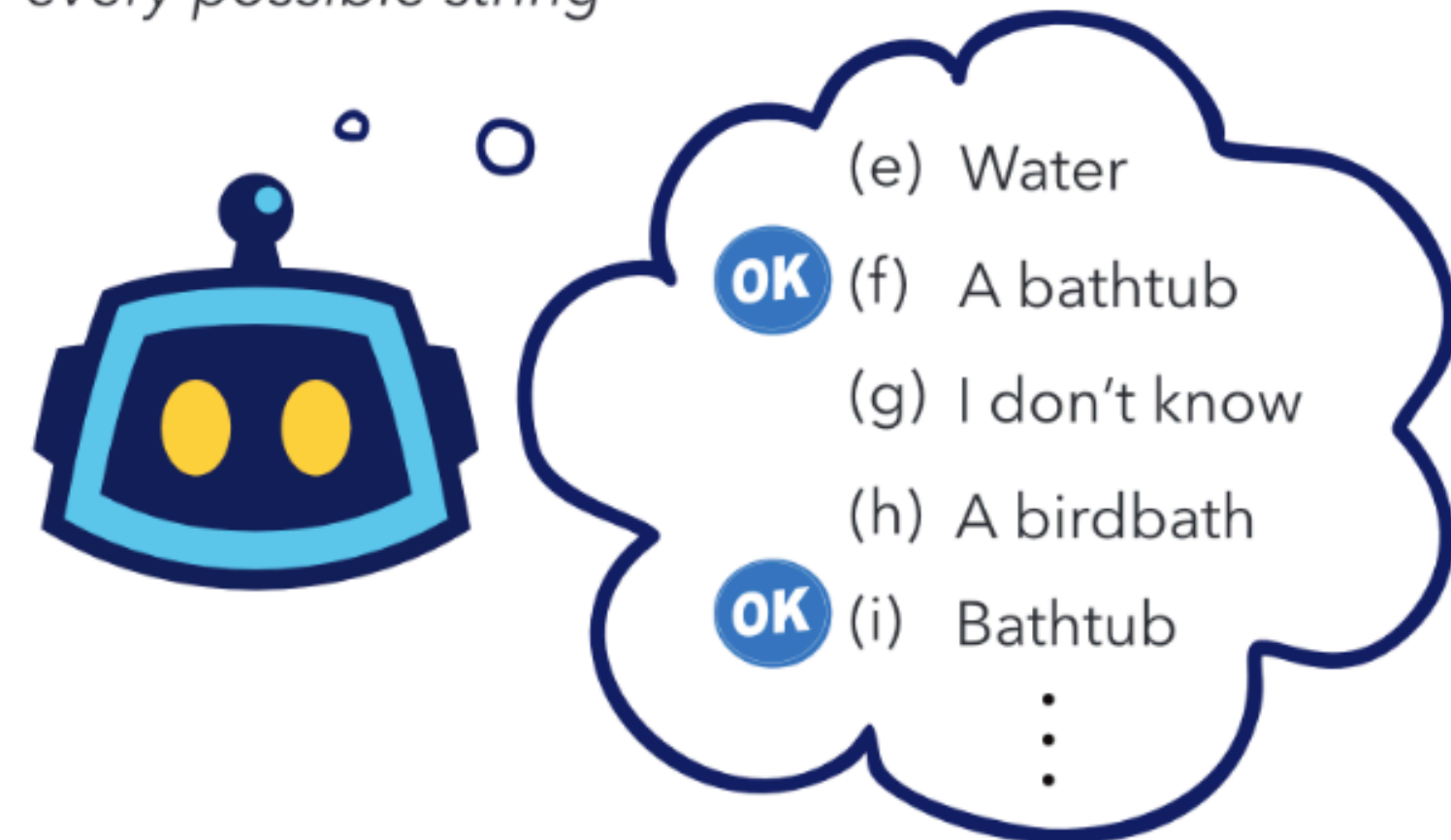**VQA models answer the question without looking at the image**

from Cadene et al. 2019

A human wants to submerge himself in water, what should he use?

Humans *select* options

- (a) Coffee cup ✗
- (b) Whirlpool bath ✓
- (c) Cup ✗
- (d) Puddle ✗

Language Models assign probability to *every possible string*

- (e) Water
- (f) A bathtub OK
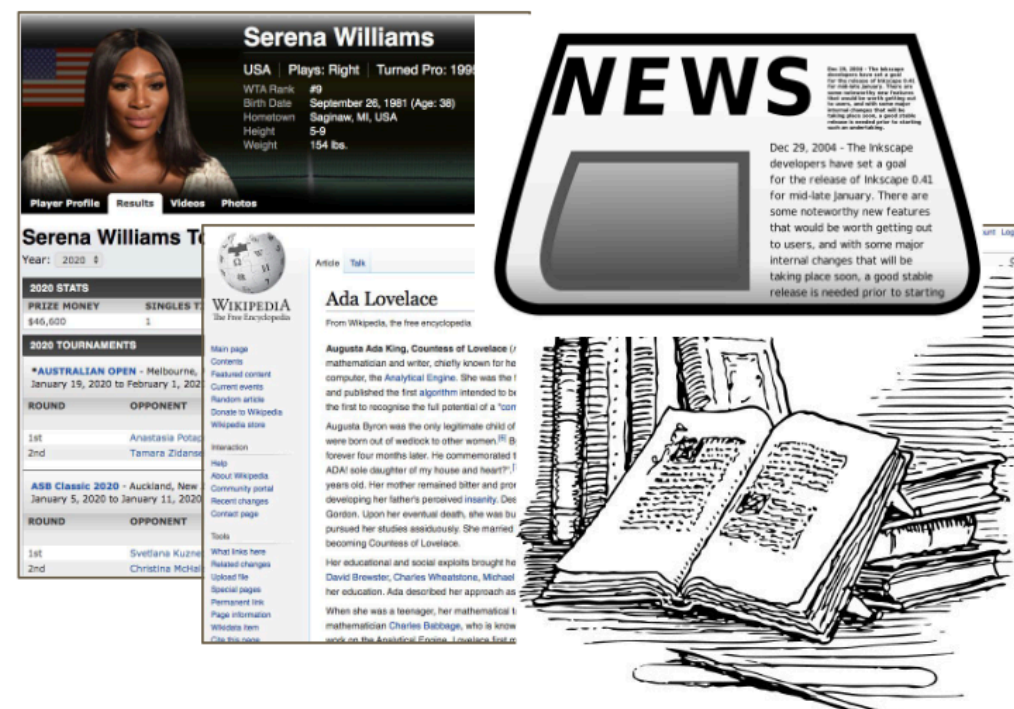- (g) I don't know
- (h) A birdbath
- (i) Bathtub OK

OK = right concept, wrong surface form

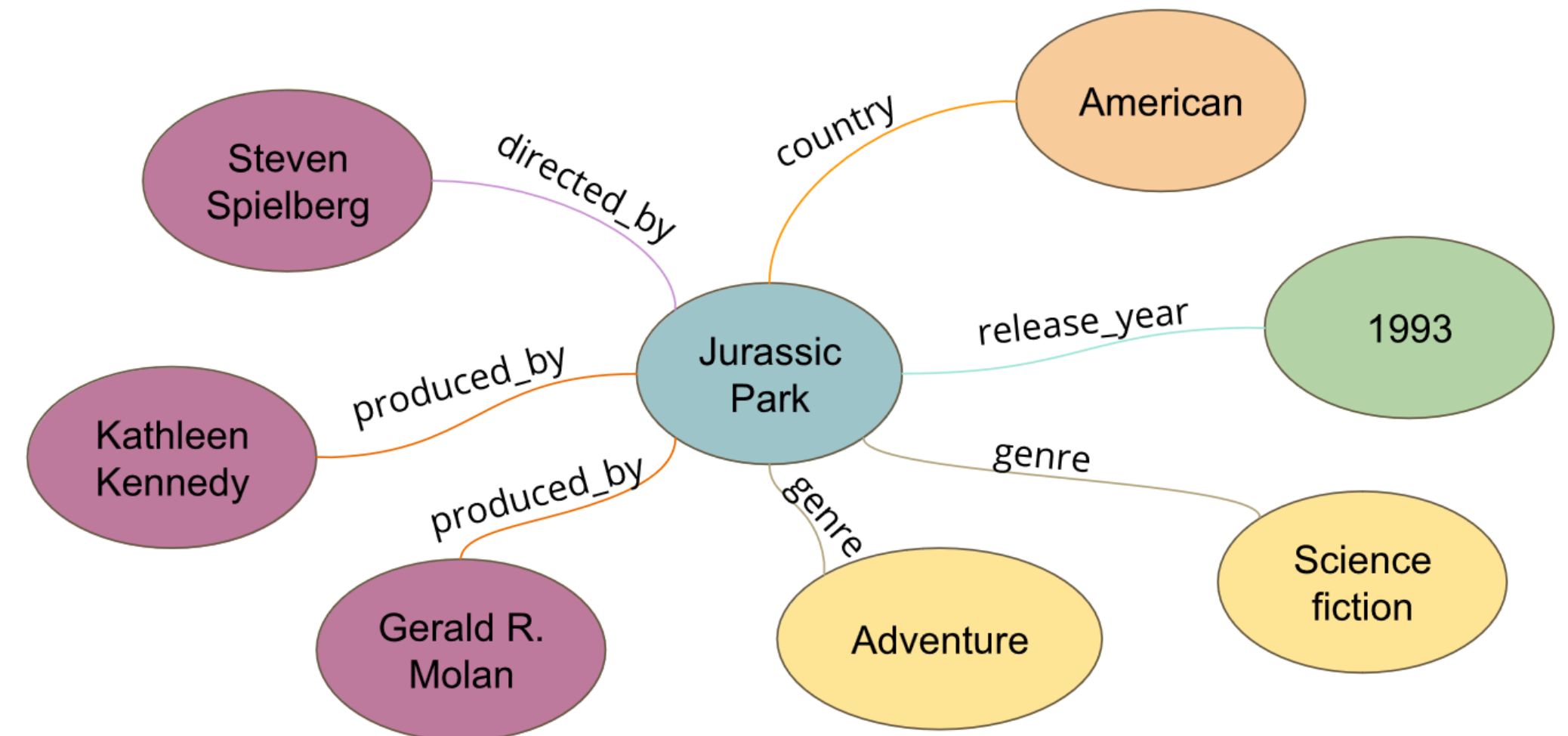from Holtzman and West et al 2022

Information Extraction

- Given a set of documents, construct some kind of structured representation of the information it encodes

(Serena Williams, born in, Florida)

Subject   Predicate (Relation)   Object

Knowledge Triple
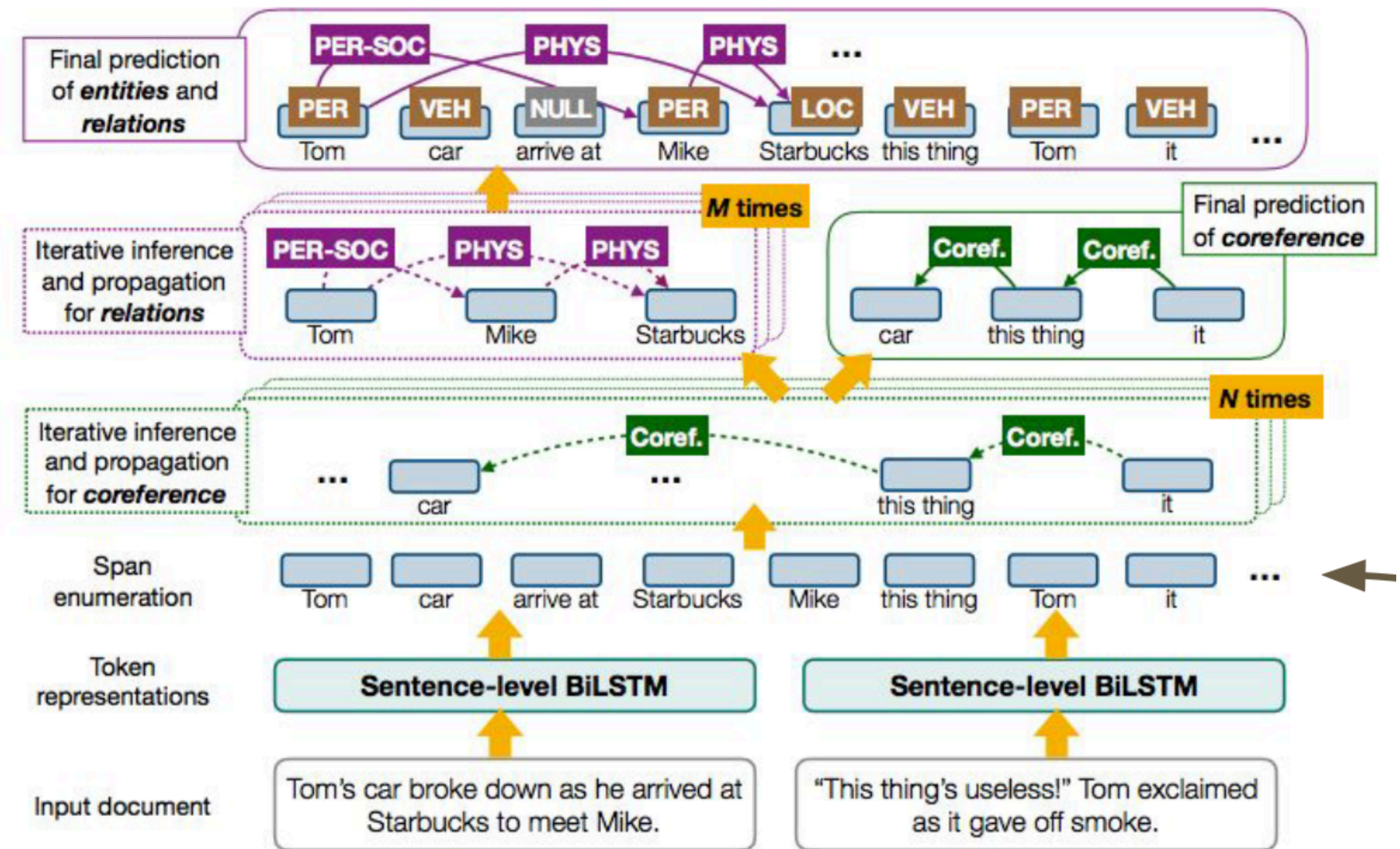
This and following slides are partially from Dong, Hajishirzi, Lockard, and Shiralkar, ACL 2020 Tutorial on multi-modal information extraction

Information Extraction Subtasks

- Comprises a number of tasks
  - Named entity recognition
  - Coreference resolution
  - Relation detection
- Evaluation is often done on these subtasks
- But this requires heavy annotation

Luan et al. 2019

Distant Supervision in Information Extraction

- Instead, mine for patterns that might express known relations in an existing knowledge base

- Evaluation can also be performed on an existing knowledge base (e.g., Freebase, https://en.wikipedia.org/wiki/Freebase_(database))

  - How well does this model recover existing relations?

  - Are the relations it recovers accurate?

Distant Supervision (Mintz et al, 2009)

Automatically generate training data using existing knowledge

Steve Jobs was the founder of Apple.

founder_of

Founder Table	
Steve Jobs	Apple
Bill Gates	Microsoft
Jeff Bezos	Amazon
...	...

Representation Learning

- High-level question: do the representations our models learn reflect reality?

- One way of measuring this: natural language inference

Following slides are mostly from Nikita Nangia, Clara Vania, and Sam Bowman (tutorial at EMNLP 2021)

Natural Language Inference
aka Recognizing Textual Entailment

Premise: *I'm watching an EMNLP talk.*

Hypothesis: *I'm having loads of fun!*

Label: {<u>entailment</u>, contradiction, neutral}

26

Dagan et al. '05, MacCartney '09, Example from MNLI

Why NLI?
NLU benchmarking and (previously) transfer learning.

- It lets you test sentence understanding comprehensively *without* grounding or semantic formalisms.

- It caught on as a benchmark task, and played a significant role in the development of self-attention and pretraining.

- It's also been useful as a *pretraining* task: Fine-tuning BERT/RoBERTa/T5/etc. on NLI data makes it easier for that model to adapt to future tasks.

  - Less clear with the latest large models.

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.

Definitely correct   Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption.   **Entailment**

Maybe correct   Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be.   **Neutral**

Definitely incorrect   Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.   **Contradiction**

Problems (optional)   If something is wrong, have a look at the FAQ, do your best above, and let us know here.

Source captions from Flickr30k: Young, et al. '14   46

Aside: Crowdsourcing

- How to reduce annotation cost and time? If you don't need experts, use cheaper labor (Snow et al. 2008)

- One very popular option: crowdsourcing platforms (mostly MTurk)

- Basic pipeline:

  - Design and pilot task (critical step!)

  - Recruit crowdworkers, e.g. via a qualification task

  - Incentive design

  - Deploying task and managing workers

Find the Answer to this Question

We believe that the answer to the question

What is Mark Twain's real name?

is contained in the below article.

Please scan the article and copy the **complete sentence** that best answers the question and paste it in the first box below. Please also identify the **answer itself** in the answer sentence and copy it in the second box below. Please copy and paste only! Do not fill the boxes by typing!

Occasionally, it might happen that you need to copy two consecutive sentences. In the *unlikely* event that the article does not contain the answer, please enter "NA" (without the quotes).

This is the article:

Twain's Account of Hanging Found

VIRGINIA CITY, Nev. (AP) -- The folklore of the Old West is often a mishmash of myth and reality, so an archivist knew he was onto something when he discovered a newspaper account of one of the state's first public hangings.

``I can see that stiff straight corpse hanging there yet," wrote the reporter, ``with its black pillow-cased head turned rigidly to one side, and the purple streaks creeping through the hands and driving the fleshy hue of life before them. Ugh!"

The reporter? Samuel Langhorne Clemens, better known as Mark Twain.

...

Please COPY AND PAST the COMPLETE ANSWER SENTENCE from the article here:

Please COPY AND PASTE (do not type) the ANSWER (usually one or a few words) from the answer sentence here:

Finished with this HIT?   Let someone else do it?
Submit HIT   Return HIT

Kaisser and Lowe 2008

29

Aside: Crowdsourcing

■ It's not trivial to do crowdsourcing well!

   ■ Well = getting high quality data

   ■ **Well = respecting workers as people**

■ Lots of work on the crowdworking ecosystem and experiences of crowdworkers, including tools they use to manage their own work (Martin et al. 2014, Irani and Silberman 2013, Kummerfeld 2021)

Find the Answer to this Question

We believe that the answer to the question

What is Mark Twain's real name?

is contained in the below article.

Please scan the article and copy the **complete sentence** that best answers the question and paste it in the first box below. Please also identify the **answer itself** in the answer sentence and copy it in the second box below. Please copy and paste only! Do not fill the boxes by typing!

Occasionally, it might happen that you need to copy two consecutive sentences. In the *unlikely* event that the article does not contain the answer, please enter "NA" (without the quotes).

This is the article:

Twain's Account of Hanging Found

VIRGINIA CITY, Nev. (AP) -- The folklore of the Old West is often a mishmash of myth and reality, so an archivist knew he was onto something when he discovered a newspaper account of one of the state's first public hangings.

``I can see that stiff straight corpse hanging there yet," wrote the reporter, ``with its black pillow-cased head turned rigidly to one side, and the purple streaks creeping through the hands and driving the fleshy hue of life before them. Ugh!"

The reporter? Samuel Langhorne Clemens, better known as Mark Twain.

...

Please COPY AND PAST the COMPLETE ANSWER SENTENCE from the article here:

Please COPY AND PASTE (do not type) the ANSWER (usually one or a few words) from the answer sentence here:

Finished with this HIT?   Let someone else do it?
Submit HIT   Return HIT

Kaisser and Lowe 2008

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.

Definitely correct   Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption.   **Entailment**

Maybe correct   Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be.   **Neutral**

Definitely incorrect   Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.   **Contradiction**

Problems (optional)   If something is wrong, have a look at the FAQ, do your best above, and let us know here.

31

Source captions from Flickr30k: Young, et al. '14   46

Premise: *A man in a blue shirt standing in front of a garage-like structure painted with geometric designs.*

Hypothesis: *A man is repainting a garage*

Label: Neutral

Premise: *Two women are embracing while holding to go packages.*

Hypothesis: *Two woman are holding packages.*

Label: Entailment

Summary

- Evaluation metric: simple classification accuracy!

- Data source: image captions

- Annotation method

  - Not a complex task — don't need to train experts or develop a complex annotation scheme

  - Instead: hire crowdworkers (on MTurk)

- However, you need to be careful…

Annotation Artifacts

For SNLI:

P: ???

H: *Someone is **not** crossing the road.*

Label: entailment, contradiction, neutral?

Poliak et al. '18, Tsuchiya '18, Gururangan et al. '18

Annotation Artifacts

Models can do moderately well on NLI datasets without looking at the premise!

Single-genre SNLI especially vulnerable. SciTail not immune, despite using no crowdworker writing.

35

Poliak et al. '18, Tsuchiya '18, Gururangan et al. '18

Instructions

The Stanford University NLP Group is collecting data for use in research on computer understanding of English. We appreciate your help!

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo.
- Write one alternate caption that **might be** a **true** description of the photo.
- Write one alternate caption that is **definitely** a **false** description of the photo.

Photo caption An older man in gray khakis walks with a young boy in a green shirt along the edge of a fountain in a park.

Definitely correct Example: For the caption *"Two dogs are running through a field."* you could write *"There are animals outdoors."*

Write a sentence that follows from the given caption.   **Entailment**

Maybe correct Example: For the caption *"Two dogs are running through a field."* you could write *"Some puppies are running to catch a stick."*

Write a sentence which may be true given the caption, and may not be.   **Neutral**

Definitely incorrect Example: For the caption *"Two dogs are running through a field."* you could write *"The pets are sitting on a couch."* This is different from the *maybe correct* category because it's impossible for the dogs to be both running and sitting.

Write a sentence which contradicts the caption.   **Contradiction**

Problems (optional) If something is wrong, have a look at the FAQ, do your best above, and let us know here.

36

Source captions from Flickr30k: Young, et al. '14 46

Aside: Benchmarking at Scale

- How well does my model perform at a wide variety of tasks?

- Collections of benchmark tasks and datasets:
    - GLUE (Wang et al. 2019, ICLR) and SuperGLUE (Wang et al. 2019, NeurIPS)
    - Dynabench (dynamic benchmarking, Kiela et al. 2021)
    - BIG-Bench (Google, 2023)
    - SuperNatural Instructions (1600+ tasks, Wang et al. 2022)

- Leaderboarding

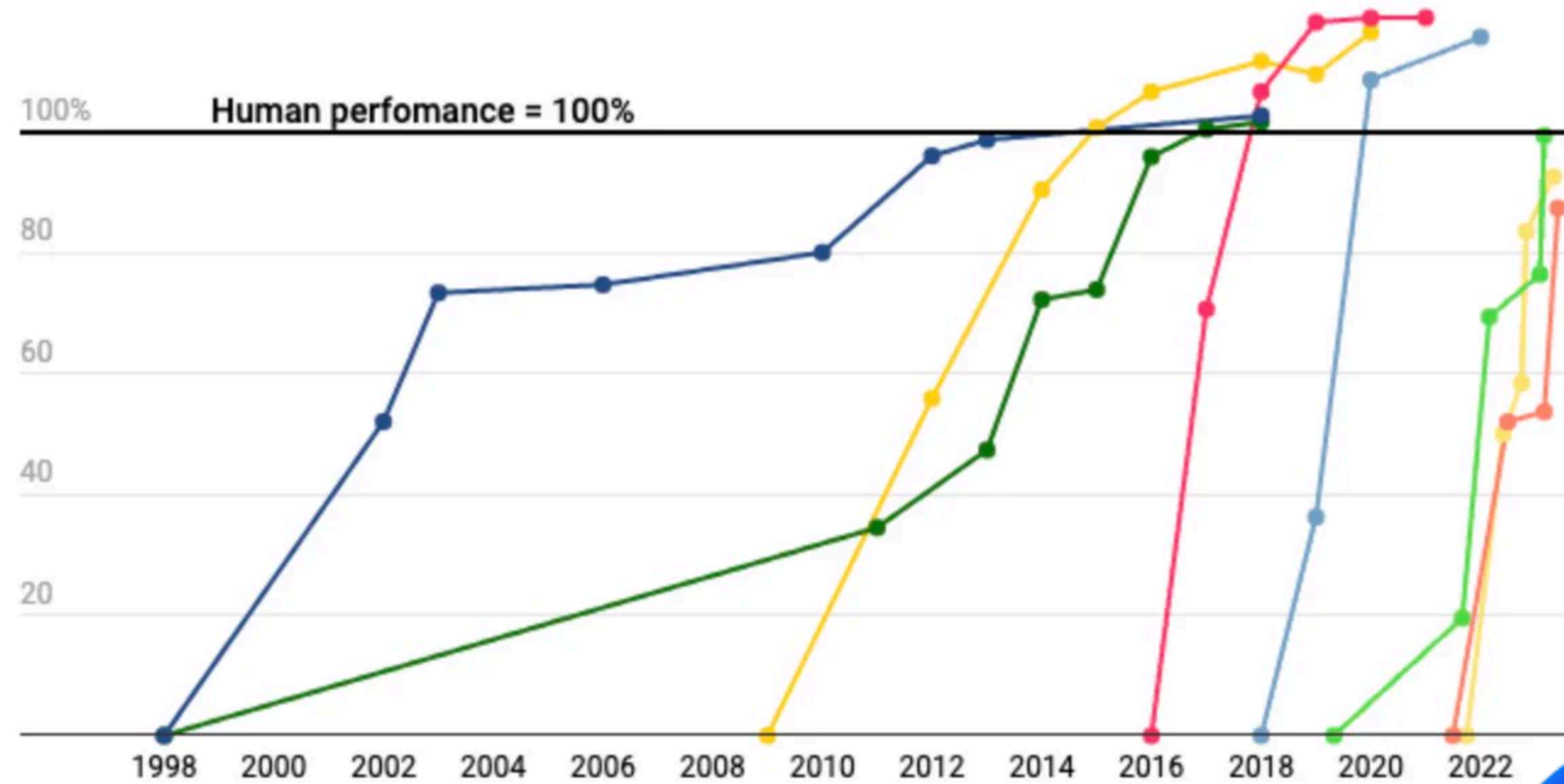SuperNatural Instructions,
Wang et al. 2022

Human Performance?

AI has surpassed humans at a number of tasks and the rate at which humans are being surpassed at new tasks is increasing

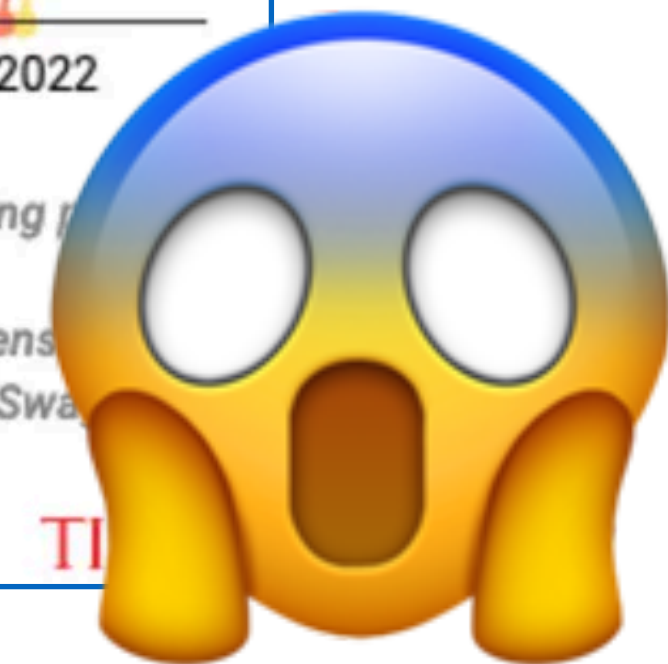State-of-the-art AI performance on benchmarks, relative to human performance

- ● Handwriting recognition ● Speech recognition ● Image recognition ● Reading comprehension
- ● Language understanding ● Common sense completion ● Grade school math ● Code generation

Human perfomance = 100%

100%
80
60
40
20

1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020 2022

For each benchmark, the maximally performing baseline reported in the benchmark paper is taken as the "starting which is set at 0%. Human performance number is set at 100%. Handwriting recognition = MNIST, Language understanding = GLUE, Image recognition = ImageNet, Reading comprehension = SQuAD 1.1, Reading comprehens SQuAD 2.0, Speech recognition = Switchboard, Grade school math = GSK8k, Common sense completion = HellaSwa Code generation = HumanEval.

Chart: Will Henshall for TIME • Source: ContextualAI

Rank	Na...		CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI
1	M...		73.3	97.5	94.2/92.3	93.5/93.1	76.4/90.9	92.5	92.1	96.7
2	J...		73.8	97.9	94.5/92.6	93.5/93.1	76.7/91.1	92.1	91.9	96.7
3	M...		72.6	97.6	93.8/91.7	93.7/93.3	76.4/91.1	92.6	92.4	97.9
4	D...		74.7	97.6	93.3/91.1	93.4/93.1	76.5/91.0	92.1	91.8	96.7
23	G...	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0		92.8	91.2

SuperGlue's leaderboard
(Wang et al 2019)

Human Performance?

- AGI is superhuman???

- Wait — what does this even mean?

- How is human performance computed?

  - Are we paying workers enough?

  - Are we training workers to complete complex tasks?

  - Are we looking into disagreements, throwing them out, or combining them in some arbitrary way (majority voting)?

  - Are our tasks too subjective?

  - Are models taking advantage of spurious correlations?

Tedeschi et al. 2023

What's missing in evaluation?

- The easier a task is to evaluate, the easier it is for a model to get the label correct with the "wrong" reasoning by taking advantage of spurious correlations

  - Maybe we shouldn't rely on automatic evaluation...

- Generalization to non-IID cases, e.g., unseen domains, languages, or tasks (Linzen 2020)

- No notion of meaningfully modeling disagreement among annotators

- No expectation of explainability in model predictions
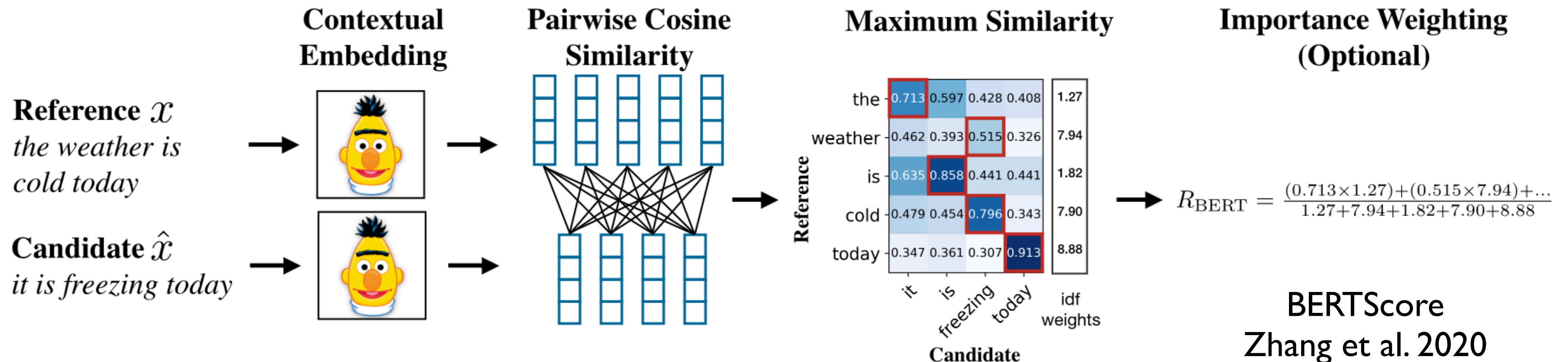
Tedeschi et al. 2023

Text Generation

- Moving beyond simple classification tasks

- Examples:

  - Summarization of structured and unstructured data, paraphrasing, text simplification

  - Creative tasks — e.g., story generation

Automatic Metrics

- If you have reference documents: can use automated metrics like BLEU, METEOR, ROUGE

- Even then, these metrics are limiting: n-gram overlap can be too strict

- Recently: neural-based evaluation metrics that allow more flexibility, e.g., BERTScore, CLIPScore (for image tasks)

BERTScore
Zhang et al. 2020

Evaluating Evaluators

- Before getting too lost in optimizing an automatic metric… how well does it reflect reality?

- Often compute correlations with (pairwise) human judgments

Metric	en↔cs	en↔de	en↔et	en↔fi	en↔ru	en↔tr	en↔zh
BLEU	.956/.993	.969/**.977**	**.981**/.971	.962/.958	.972/.977	.586/.796	.968/.941
ITER	.966/.865	.990/.978	.975/**.982**	.989/.966	.943/.965	.742/.872	.978/ –
RUSE	.974/ –	.996/ –	.988/ –	**.983**/ –	.982/ –	.780/ –	.973/ –
YiSi-1	.942/.985	.991/.983	.976/.976	.964/.938	**.985/.989**	**.881/.942**	.943/.957
P_{BERT}	.965/.989	.995/.983	**.990/.970**	.976/.951	.976/.988	.846/.936	.975/.950
R_{BERT}	**.989/.995**	.997/**.991**	.982/**.979**	.989/**.977**	**.988/.989**	.540/**.872**	**.981/.980**
F_{BERT}	.978/**.993**	.998/.988	.989/.978	.983/.969	.985/.989	.760/.910	**.981**/.969
F_{BERT} (idf)	.982/.995	**.998**/.988	**.988**/.979	**.989**/.969	.983/.987	.453/.877	.980/.963

BERTScore
Zhang et al. 2020

Human Judgments

- Pairwise judgments: how often do humans prefer reference versus generated text?

- Requires crowdsourcing

  - If we crowdsource evaluation every time, we've lost the ability to perform exact comparisons between models

- Requires a reference text

  - If there's no gold standard reference, often just a comparison between a baseline and a proposed method

- **Story generation: evaluation is subjective**

1. **Interesting.** Interesting to the reader.
2. **Coherent.** Plot-coherent.
3. **Relevant.** Faithful to the initial premise.
4. **Humanlike.** Judged to be human-written.

We additionally track how often generated stories suffer from any of the following writing issues:

1. *Narration.* Jarring change(s) in narration and/or style.
2. *Inconsistent.* Factually inconsistent or containing very odd details.
3. *Confusing.* Confusing or difficult to follow.
4. *Repetitive.* Highly repetitive.
5. *Disfluent.* Frequent grammatical errors.

RE3, Yang et al. 2022

Questions:

- 1) Which story do you prefer / find more interesting overall?

  ○ Story A
  ○ Story B
  ○ Both are about equally good
  ○ Neither is good

- 2) Which story has a more coherent overarching plot?

  ○ Story A
  ○ Story B
  ○ Both are about equally good
  ○ Neither is good

- 3) Which story's plot is closer to the premise?

  ○ Story A
  ○ Story B
  ○ Both are about equally good
  ○ Neither is good

- 4) Indicate which of the following problems are present in Story A (possibly none, possibly more than one).

  ☐ Jarring change(s) in narration or style

  ☐ Factual inconsistencies/oddities

  ☐ Very confusing or hard to understand

  ☐ Often ungrammatical or disfluent

  ☐ Highly repetitive

  ☐ None of the above

- 5) Indicate which of the following problems are present in Story B (possibly none, possibly more than one).

  ☐ Jarring change(s) in narration or style

  ☐ Factual inconsistencies/oddities

  ☐ Very confusing or hard to understand

  ☐ Often ungrammatical or disfluent

  ☐ Highly repetitive

  ☐ None of the above

- 6) Do you think Story A was written by a human?

  ○ Yes
  ○ No

- 7) Do you think Story B was written by a human?
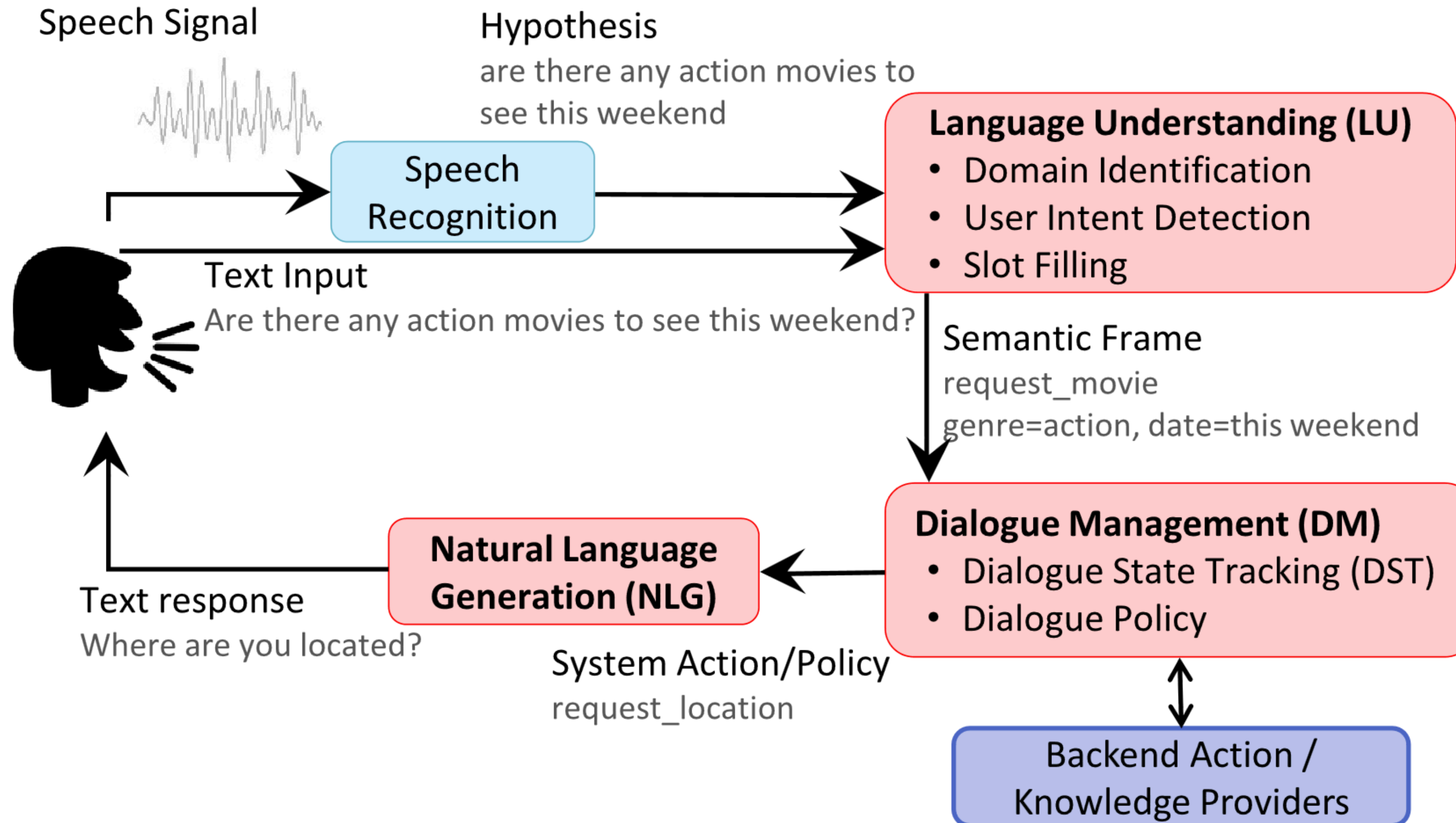
  ○ Yes
  ○ No

Dialogue and Interactive Systems

- What is language?

- Fundamentally: an interactive tool to get things done in the world

- How well do our systems *use* language?

Following slides are mostly from Yun-Nung Chen, Asli Celikyilmaz, and Dilek Hakkani-Tür, from ACL 2017 tutorial on deep learning for dialogue systems, and Campagna and Lam from CS 224v

Dialogue Systems

Speech Signal

Hypothesis
are there any action movies to
see this weekend

Speech
Recognition

Language Understanding (LU)
- Domain Identification
- User Intent Detection
- Slot Filling

Text Input
Are there any action movies to see this weekend?

Semantic Frame
request_movie
genre=action, date=this weekend

Natural Language Generation (NLG)

Dialogue Management (DM)
- Dialogue State Tracking (DST)
- Dialogue Policy

Text response
Where are you located?

System Action/Policy
request_location

Backend Action /
Knowledge Providers

Evaluation

- **Standardized benchmark / shared task: Dialogue State Tracking Challenge**

  - **Slot-filling over time**

  - **Requires designing frame representations of dialogue state**

Sentences That Cannot Be Represented As Slots

"I was hoping you could **recommend** something".

"Are there any churches **or** museums on the east side?"

"I would like the **latest** train leaving that will arrive by 9:15 please".

A: Hello! This is Concierge Service. I can help you find attractions, hotels, restaurants in Cambridge.

U: I'm looking for a restaurant.
[Domain=Restaurant]

A: What cuisine would you like?

U: I would like Italian food.
[Domain=Restaurant, Food=Italian]

A: Would you like a cheap, moderate or expensive Italian restaurant?

U: Actually, never mind, let's do Chinese.
[Domain=Restaurant, Food=Chinese]

A: Would you like a cheap, moderate or expensive Chinese restaurant?

Evaluation

- More subjective metrics

  - Human evaluation of a dialogue they observe

  - Doesn't measure how dialogue system might be used in practice

  - E.g., won't model how errors affect later parts of the dialogue

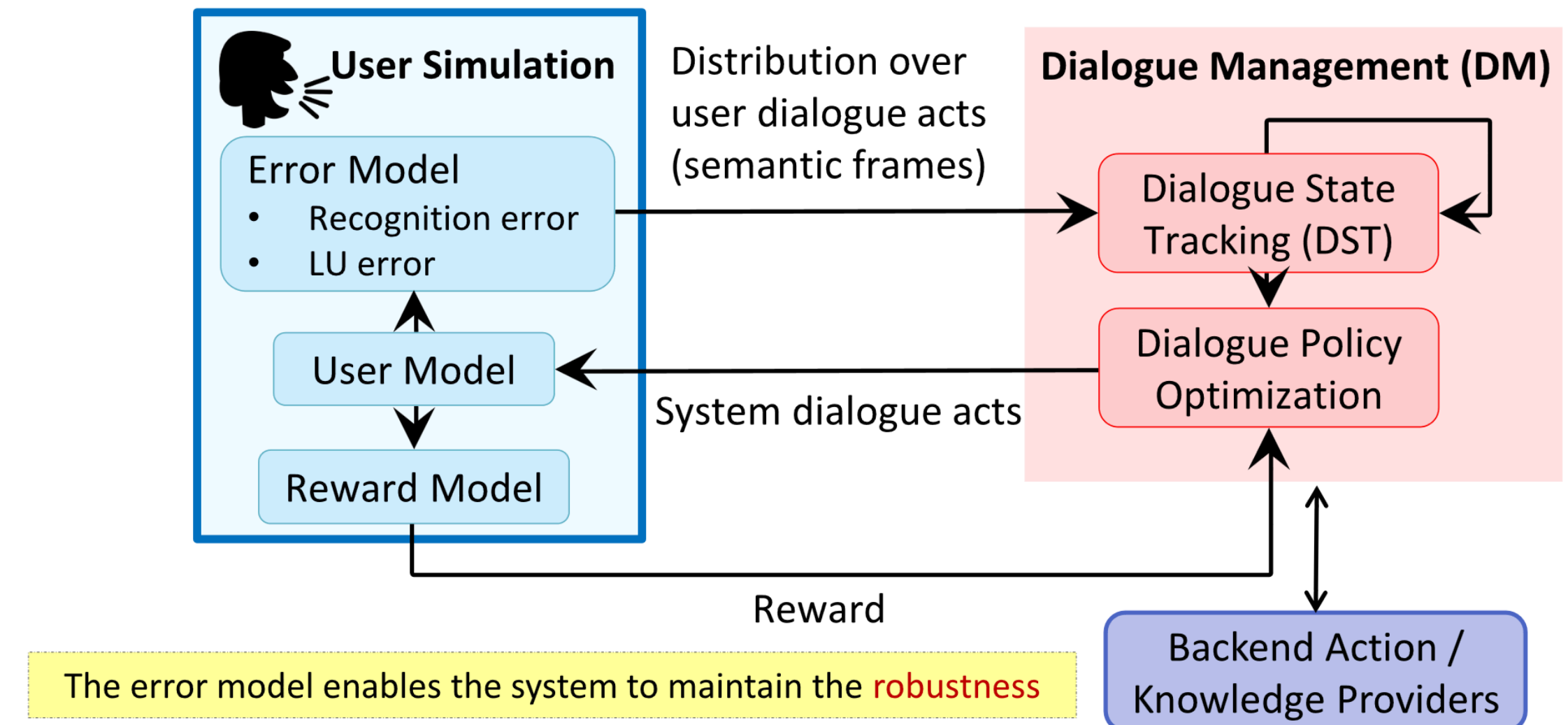| Q1	Do you think you understand from the dialog what the user wanted?
Opt	1) No clue 2) A little bit 3) Somewhat 4) Mostly 5) Entirely
Aim	*elicit the Worker's confidence in his/her ratings.*
Q2	Do you think the system is successful in providing the information that the user wanted?
Opt	1) Entirely unsuccessful 2) Mostly unsuccessful 3) Half successful/unsuccessful 4) Mostly successful 5) Entirely successful
Aim	*elicit the Worker's perception of whether the dialog has fulfilled the informational goal of the user.*
Q3	Does the system work the way you expect it?
Opt	1) Not at all 2) Barely 3) Somewhat 4) Almost 5) Completely
Aim	*elicit the Worker's impression of whether the dialog flow suits general expectations.*
Q4	Overall, do you think that this is a good system?
Opt	1) Very poor 2) Poor 3) Fair 4) Good 5) Very good
Aim	*elicit the Worker's overall impression of the SDS.*
Q5	What category do you think the dialog belongs to?
Opt	1) Task is incomplete 2) Out of scope 3) Task is complete
Aim	*elicit the Worker's impression of whether the dialog reflects task completion.*

Yang et al. 2012

Evaluation

- "Online" evaluation: user simulators

- Allows scaling up experiments

- Also is stable across systems

  - But does not reflect real-world complexity of actual use cases

  - E.g., users adapt to systems over interactions

User Simulation

Error Model
- Recognition error
- LU error

User Model

Reward Model

Distribution over user dialogue acts (semantic frames)

System dialogue acts

Reward

Dialogue Management (DM)

Dialogue State Tracking (DST)

Dialogue Policy Optimization

Backend Action / Knowledge Providers

The error model enables the system to maintain the robustness

Wizard-of-Oz (Kelley 1984)

Help Desk: Hello, welcome to the TownInfo centre. I can help you find a restaurant or hotel, look for tourist information, book a train or taxi. How may I help you ?

Customer : I want a place to stay in the east.

Help Desk : I have 6 guesthouses and 1 hotel on the east side. What's your price range?

Customer : Doesn't matter too much. I'd like a 4 star property, though, and would prefer one of the guesthouses.

Help Desk : I'd recommend 517a coldham lane. Would you like me to book a room?

Customer : Could you give me their phone number? I would like to verify that they have free parking.

Help Desk : Allenbell does have parking and the phone is 01223210353. Can I help with anything else?

[Next turn]

Customer : (**Your response**)

you need to go through the dialogue first by clicking the 'next turn' button

30

What topics were mentioned in **this turn**:

General: ☐ Booking: ☐ Restaurant: ☐ Hotels: ☐ Attraction: ☐

Hospital: ☐ Police: ☐ Train: ☐ Taxi: ☐ Bus: ☐

[Submit the HIT]

Please try to chat about the following topic:

Task MUL0002:

- You are traveling to and looking forward to try local restaurants.
- You are looking for a **place to stay**. The hotel should be in the **east** and should **include free parking**.
- The hotel should have **a star of 4** and should be in the type of **guesthouse**.
- Make sure you get **address** and **phone number**.
- You are also looking for a **place to dine**. The restaurant should be in the **moderate** price range and should serve **australian** food.
- If there is no such restaurant, how about one that serves **turkish** food.
- Once you find the **restaurant** you want to book a table for **4 people** at **17:45** on **friday**.
- Make sure you get the **reference number**

--- The End ---

MultiWOZ, Budzianowski et al. 2018

Wizard-of-Oz (Kelley 1984)

Task MUL0002

Help Desk: Hello, welcome to the   TownInfo centre. I can help you find a restaurant or hotel, look for tourist information, book a train or taxi. How may I help you ?

Customer :   I want a place to stay in the east.

Help Desk :   I have 6 guesthouses and 1 hotel on the east side. What's your price range?

Customer :   Doesn't matter too much. I'd like a 4 star property, though, and would prefer one of the guesthouses.

Help Desk :   I'd recommend   517a coldham lane. Would you like me to book a room?

Customer :   Could you give me their phone number? I would like to verify that they have free parking.

`Next turn`

| Restaurant | Hotel | Attraction | Hospital | Police | Train | Taxi | Bus |

Please **modifiy** the following answers based on the latest customer response:

- What does the user want?

Is the user looking for a specific hotel **by name**?	not mentioned
What is the **hotel type** the user wants?	guesthouse
What is the **area** the user wants?	east
What is the **price range** the user wants?	not mentioned
What is the **star of the hotel** the user wants?	4
Does the user need **internet** ?	not mentioned
Does the user need **parking** ?	not mentioned

`Lookup`

Help Desk : **(Your response)**

> you need to fill in the questionnaires above first.
>
> 30

☐ end-of-dialogue?

`Submit the HIT`

MultiWOZ, Budzianowski et al. 2018

52

Interactive Systems

- **What is language?**

- **Fundamentally: an interactive tool to get things done in the world**

- **How well do our systems *use* language?**

Following slides are from Alane Suhr and Yoav Artzi,
EMNLP 2021 tutorial on crowdsourcing

CerealBar

A situated collaborative game with
sequential natural language instruction

CerealBar

- **Interaction:** participants respond to each others' language and behavior across multiple turns

- **Collaboration:** participants are incentivized to coordinate using language

- **Key difference from existing interactive systems:** evaluate success of language use via measuring collaboration success!

Game Design

Environment

- Passable **terrain**

- **Obstacles** to navigate around (terrain and landmarks)
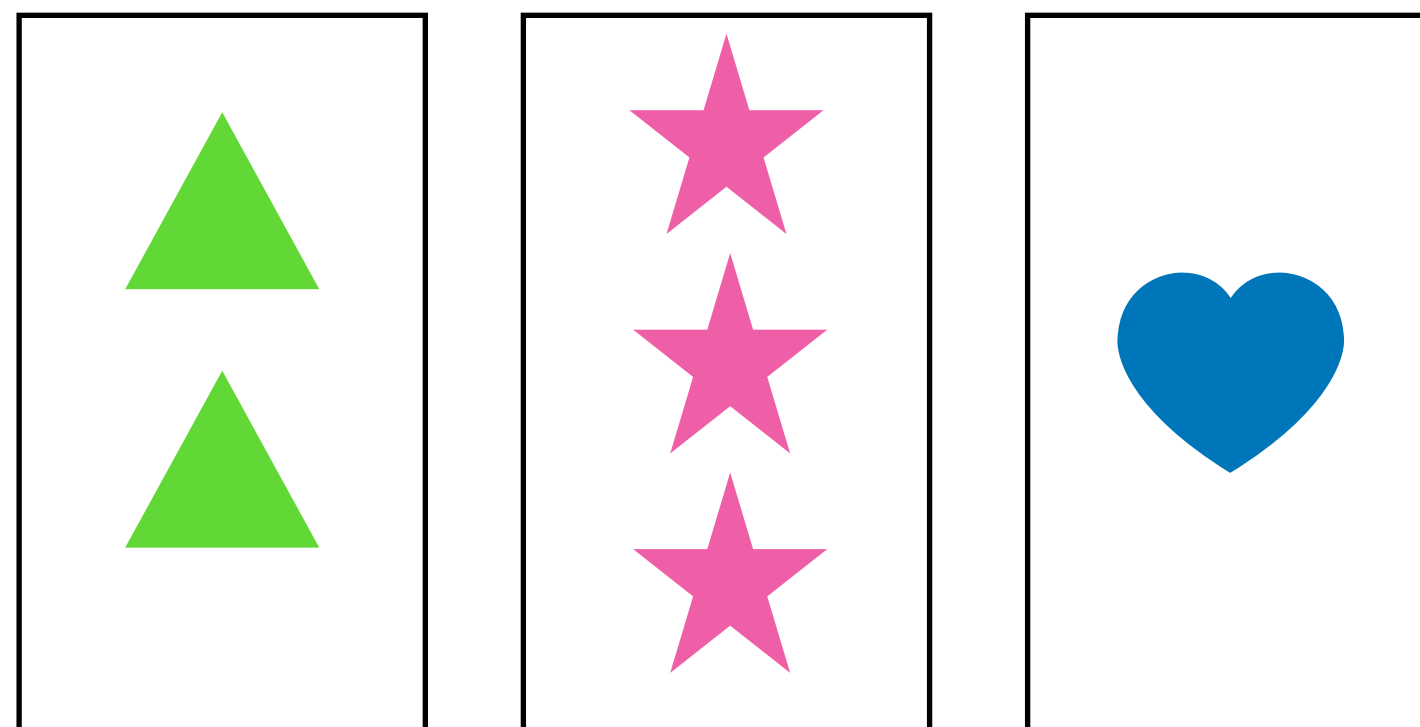
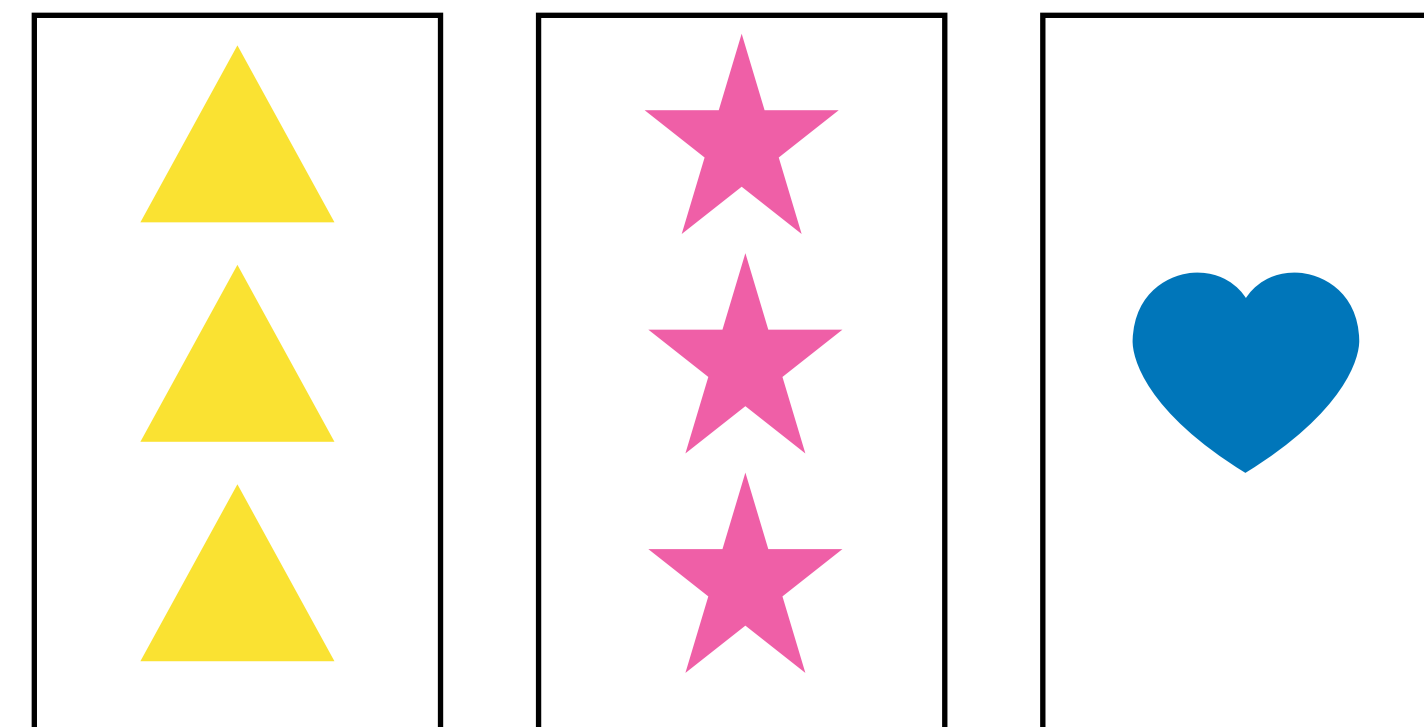- **Cards** can be selected or unselected

Collaboration

Collaboration

- Collect valid sets of three cards

- Valid: unique color, shape, and count

- Each set completed is one point

- Goal: maximize game score

✔ **Valid Set**

❌ **Invalid Set**

(two cards with three objects)

Collaboration

Leader

Follower

Language

- Since players are working on the same set together, they need to coordinate their actions

- Solution to this: communicate!

- To make it easier for us to build systems that play this game, we use unidirectional communication

Instruction

- **Leader's role:** give instructions to the follower

  - Allow flexibility in instruction giving: write as many instructions as they want per turn, as long as the follower has one to follow

- **Follower's role:** follow the instructions

  - Also flexible: follow as many instructions as they want per turn, or take multiple turns for an instruction

Incentivizing Instruction

- Players have different abilities and knowledge, and must use language to bridge those differences

- **Observability**: leader sees the whole board, but follower only sees a first-person view

  - Leader is responsible for planning what cards both players should get

  - Follower is disincentivzed to wander off or select unmentioned cards

  - Leader's instructions need to be grounded in the follower's first-person view (e.g., contain spatial relations)

- **Action:** follower has 10 steps per turn, while leader has only 5

  - Encourages leader to delegate longer, more complex paths to the follower (i.e., more interesting language)
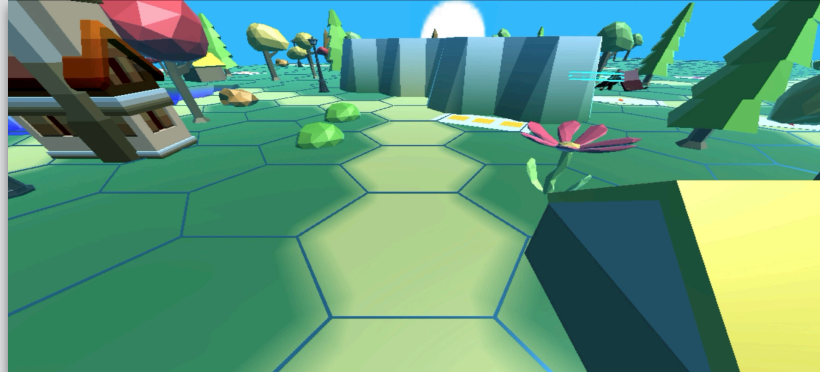
Interaction

- Fundamental to CerealBar: interaction across multiple turns

- This allows:

  - Adaptation to the other player's behavior

  - Correction of mistakes

  - Formation of common ground

Tasks Supported by CerealBar

Task I: map leader instructions to follower actions

$$f(\text{instruction},$$
$$\text{history}, \quad \text{} \quad)=\text{actions}$$

Task II: generate leader instructions

$$f(\quad \text{} \quad , \text{history})=\text{instruction}$$

Meta-Level Challenges of NLP Research

- Domain generalization, from training to test…

  - Low-resource languages

  - Specific domain applications requiring expertise

  - Real-world deployment: how do users adapt their behavior to agents they interact with?

- Replicability

- Variance and subjectivity of tasks

- Next week: panel