ECE594: Mathematical Models of Language

Spring 2022

Lecture 11: Commonsense Reasoning with NLP

Slides credit: Vered Schwartz, Maarten Sap

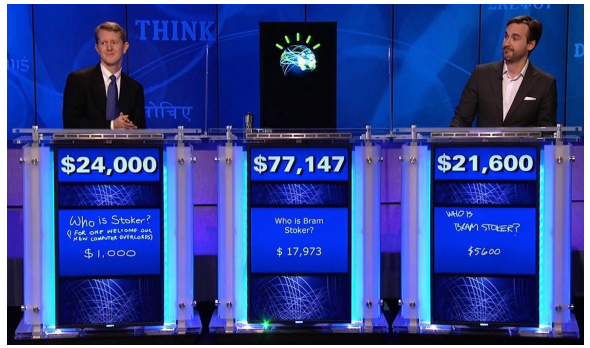
Logistics

- Project progress presentation 3/30

UNIT 2

- Low-Resource NLP
- Summarization
- Dialogue
- Question Answering
- Commonsense Reasoning

AI vs. Human Intelligence



2011

2015

Forbes

Feb 19, 2015, 01:06pm EST

Microsoft's Deep Learning Project Outperforms Humans In Image Recognition

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIBE

CADE METZ BUSINESS 09.27.2016 01:00 PM

An Infusion of AI Makes Google Translate More Powerful Than Ever

The Internet giant has unveiled an English-Chinese translation system built entirely on deep neural networks, saying it reduces error rates by 60 percent.

2016

2017

GeekWire NEWS JOBS EVENTS RESOURCES ABOUT Search

Microsoft claims new speech recognition record, achieving a super-human 5.1% error rate

BY TODD BISHOP on August 20, 2017 at 7:44 pm

New AI Model Exceeds Human Performance at Question Answering

Dave Costenaro Follow

Nov 21, 2018 · 5 min read

2018

2020

VB The Machine GamesBeat Jobs Special Issue Become a Mem

The Machine

Making sense of AI

Facebook open-sources Blender, a chatbot people say 'feels more human'

Kyle Wiggers @Kyle_L_Wiggers April 29, 2020 8:00 AM

Is AI a solved problem?

What do we know so far?

- Models can be brittle
  - Stumble on instances unlike training data

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$

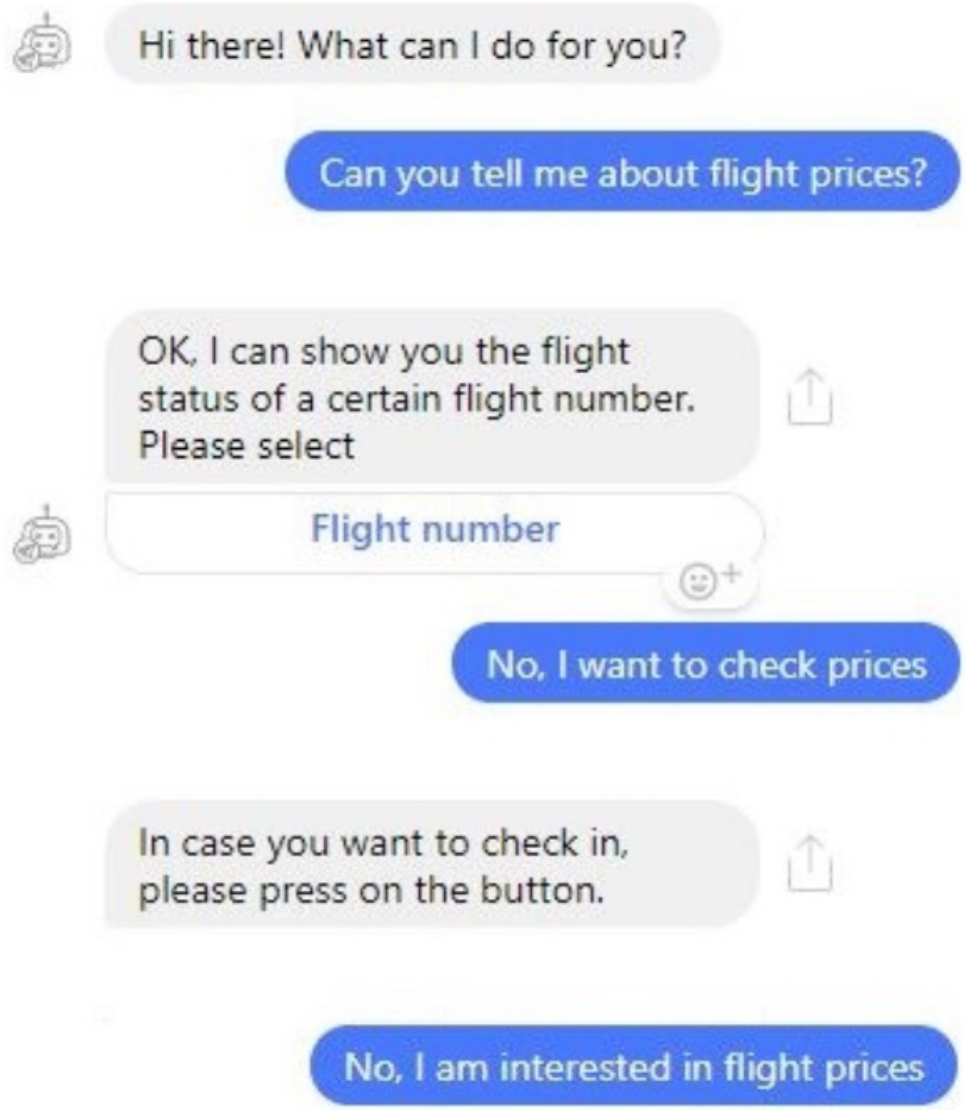
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence



Are we solving the task or fitting model to dataset?

What is Commonsense?

Merriam Webster dictionary:

Commonsense is "sound and prudent judgment based on a simple perception of the situation or facts."

What is Commonsense

- Basis of **practical knowledge** and **reasoning**
  - Concerns everyday situations and events
  - Commonly shared among most people
  - Interpretation of world around us
    - Door open or closed?

      Closet door?

      Fridge door?

What is Commonsense

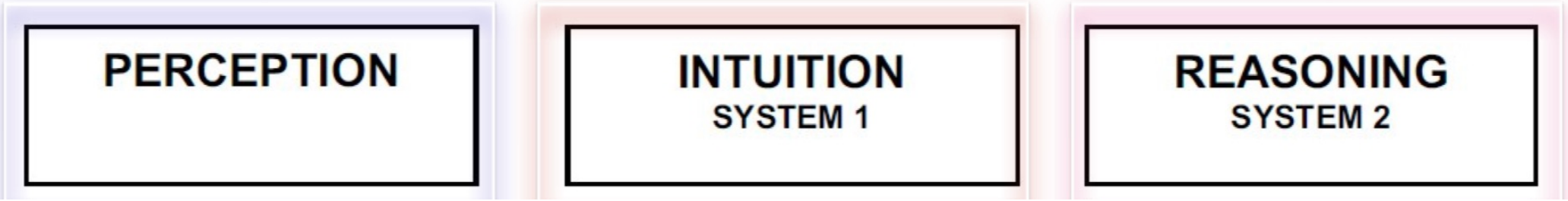
- Helps human-human interaction
  - Essential to live and interact reasonably and safely

- Helps human-machine interaction
  - Essential for AI to understand human needs and actions

Human Cognition

Kahneman's "three cognitive systems"

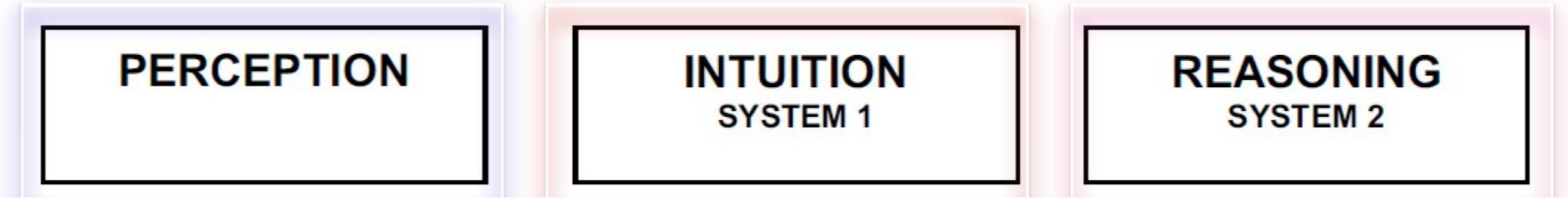
— "*Maps of Bounded Rationality: …*" (Kahneman 2003)



Where are we and where do we go

Kahneman's "three cognitive systems"

— "*Maps of Bounded Rationality: …*" (Kahneman 2003)



- Intuitive inferences on
  - pre-conditions and post-conditions
  - what happens before and after?
  - motivations and intents
  - mental and emotional states

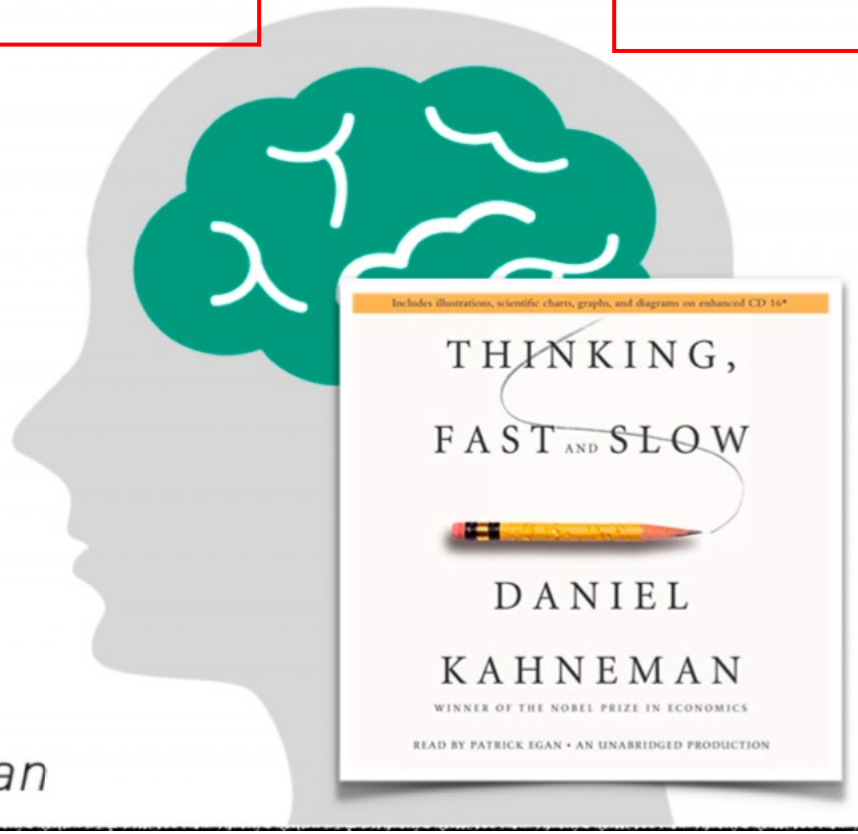
- solving puzzles
- writing programs
- proving logic theorems

Where are we and where do we go?

SYSTEM 1
Intuition & instinct

95%

Unconscious
Fast
Associative
Automatic pilot



SYSTEM 2
Rational thinking

5%

Takes effort
Slow
Logical
Lazy
Indecisive

Source: Daniel Kahneman

Where are we and where do we go

Kahneman's "three cognitive systems"

— "*Maps of Bounded Rationality: …*" (Kahneman 2003)

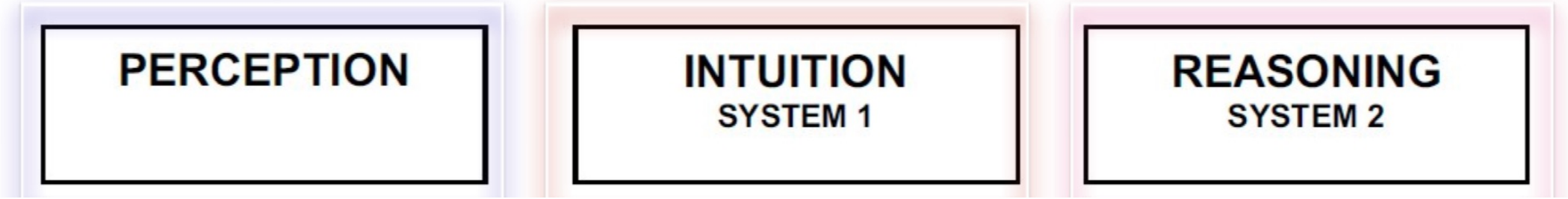


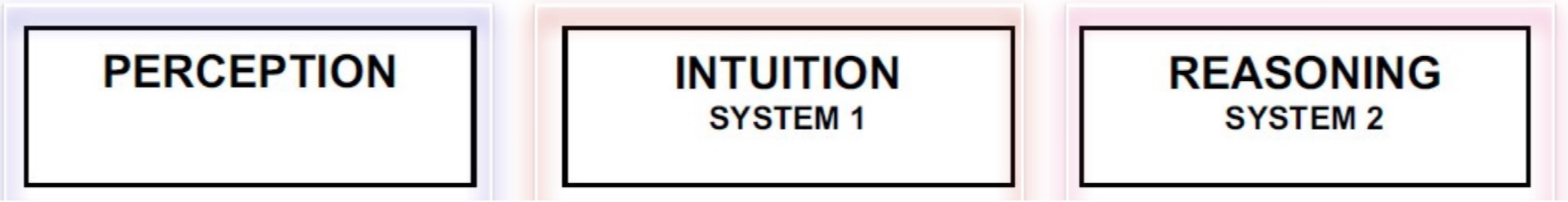
Image segmentation
Speech recognition

- Intuitive inferences on
  - pre-conditions and post-conditions
  - what happens before and after?
  - motivations and intents
  - mental and emotional states

- solving puzzles
- writing programs
- proving logic theorems

Kahneman's "three cognitive systems"

— "*Maps of Bounded Rationality: …*" (Kahneman 2003)



Language models and
Deep learning models

Processing Commonsense

- Early work in the 1980s

Position Paper on Common-sense and Formal Semantics

Geoffrey Nunberg
Xerox PARC and CSLI, Stanford

1. A philological excursus

I'm not sure what I'm doing on this panel, but I thought it would be helpful if we could start at the beginning. It's interesting to note that both the dictionary and common sense were eighteenth-century inventions. This is no coincidence; in fact, it's entirely appropriate that the most celebrated

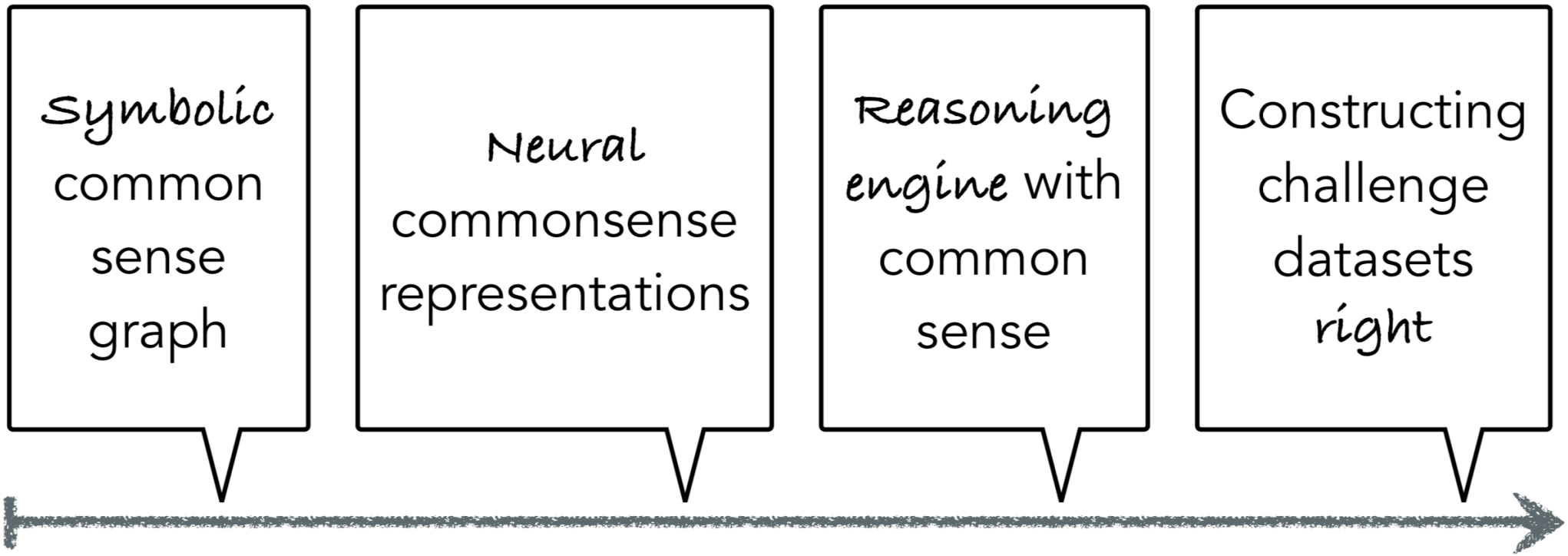
Processing Commonsense

- No concrete computational advances
  - Lack of conceptualization/representation
  - Not strong computational models/computing power
  - Not much data
  - No crowdsourcing

"Commonsense reasoning is the new frontier of artificial intelligence." Yejin Choi, UW

Path to Commonsense

- Brute force?
  - Larger and deeper networks



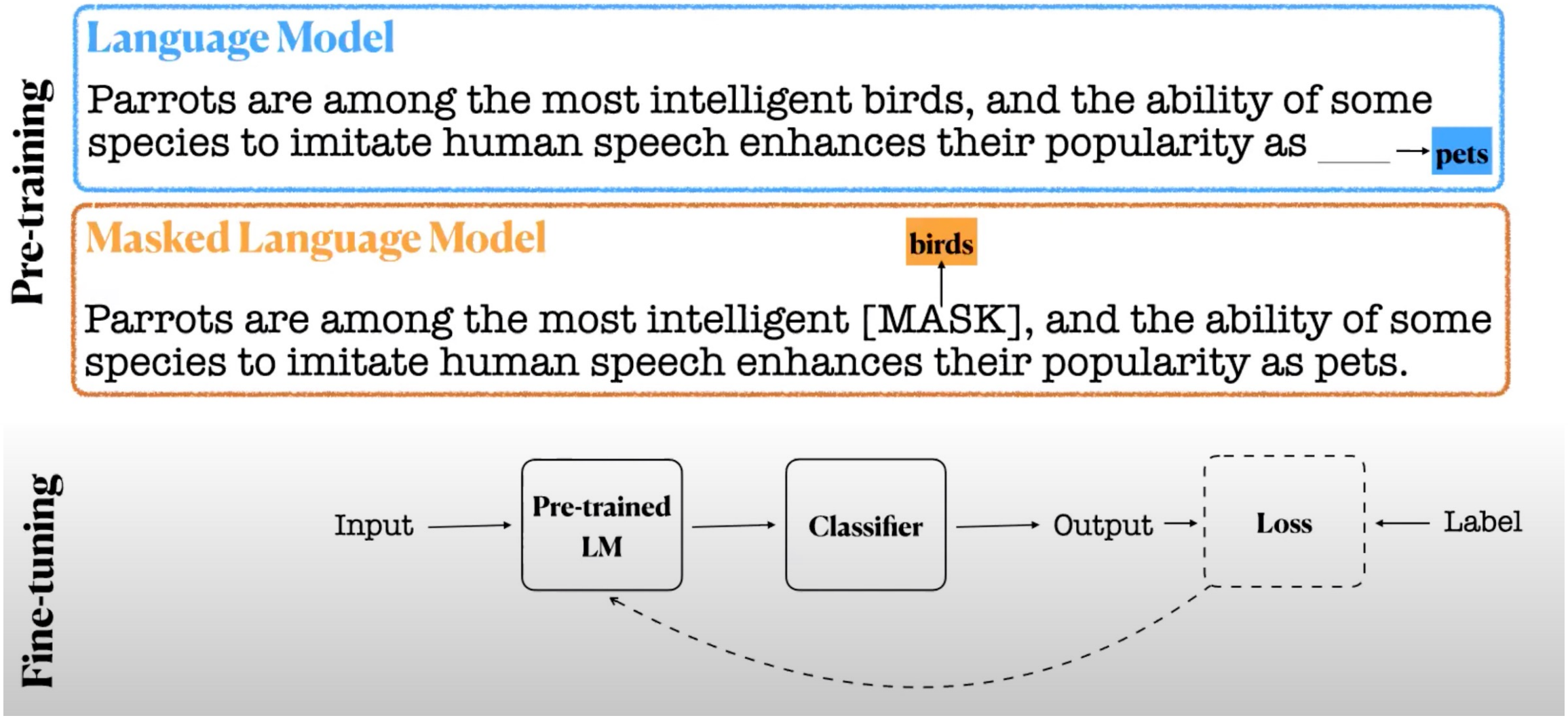
Commonsense and NLP

- Knowledge in Pre-trained Language Models
- Commonsense benchmarks
- Commonsense knowledge sources
- Endowing NN with commonsense

Knowledge in Pretrained LM

- Self-supervised models trained on large corpora
  - Trained to predict the next word in sequence or masked word in sentence

Knowledge in Pretrained LM



Knowledge in Pre-trained LM

- Do pretrained models already have commonsense?
  - What kind of commonsense knowledge do they have?

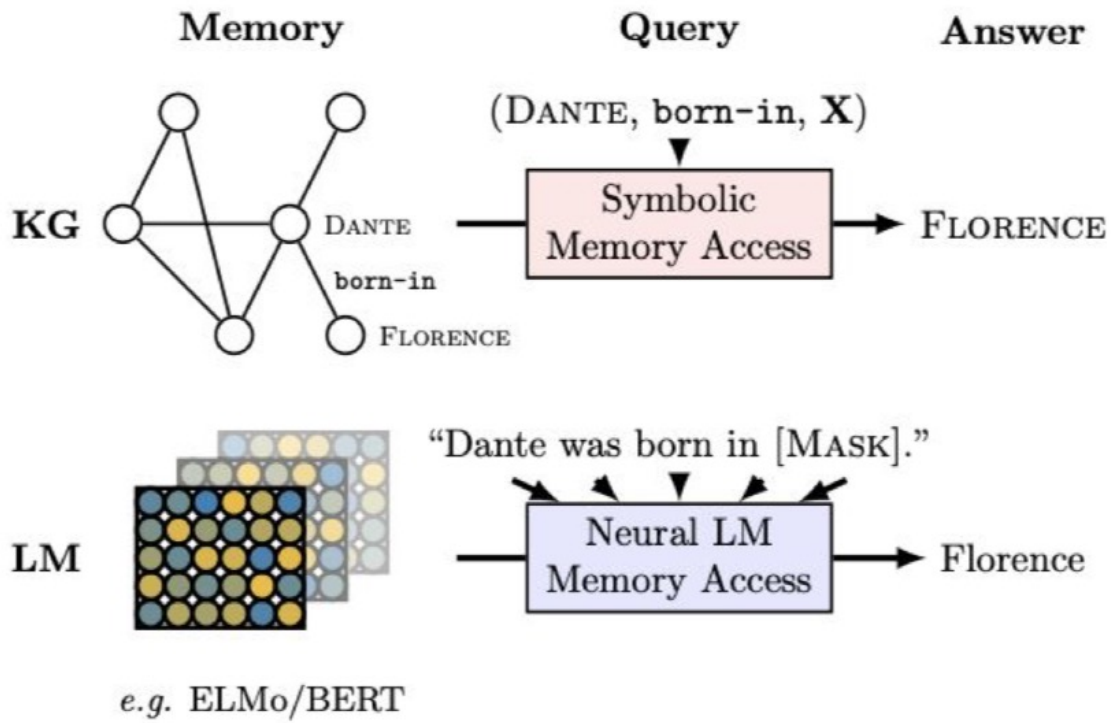
Knowledge in Pre-trained LM

- Do pretrained models already have commonsense?
- Use for knowledge-base completion
  - ConceptNet, WikiData

Knowledge-Base Completion using LM

- Task: Populate Knowledge bases
- Challenge: Need complex NLP pipelines involving entity extraction, coreference resolution, entity linking
- Approach:
  - Convert KB relations to NL templates
  - Use LMs to fill templates and score

- **Petroni et al. (2019)**:
  - LMs: ELMo / BERT
  - Templates: Hand-crafted templates
  - KBs: ConceptNet and Wikidata
  - Conclusion: BERT performs well but all models perform poorly on many-to-many relations



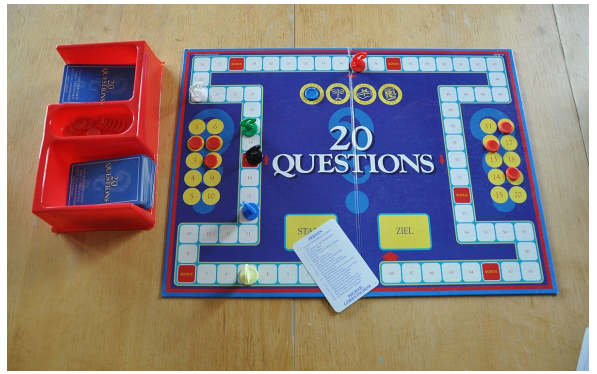
- **Feldman et al. (2019)**:
  - BERT
  - Hand-crafted templates scored by GPT2
  - ConceptNet, mining from Wikipedia
  - Performs worse than supervised methods on ConceptNet but is more likely to generalize to different domains

Candidate Sentence S_i	$\log p(S_i)$
"musician can playing musical instrument"	-5.7
"musician can be play musical instrument"	-4.9
"musician often play musical instrument"	-5.5
"a musician can play a musical instrument"	$\mathbf{-2.9}$

Table 1: Example of generating candidate sentences. Several enumerated sentences for the triple `(musician, CapableOf, play musical instrument)`. The sentence with the highest log-likelihood according to a pretrained language model is selected.

Commonsense in Pre-trained LM

- Do pretrained models already have commonsense?
- Knowledge-base completion
- Can pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?



Distinguish concepts (Weir et al. 2020)

Prompt	Model Predictions
A ___ has fur.	dog, cat, fox, ...
A ___ has fur, is big, and has claws.	cat, **bear**, lion, ...
A ___ has fur, is big, has claws, has teeth, is an animal, eats, is brown, and lives in woods.	**bear**, wolf, cat, ...

- The concept bear as a target emerging as the highest ranked predictions of neural LM
- RoBERTa > BERT

Mean Reciprocal Rank (MRR)

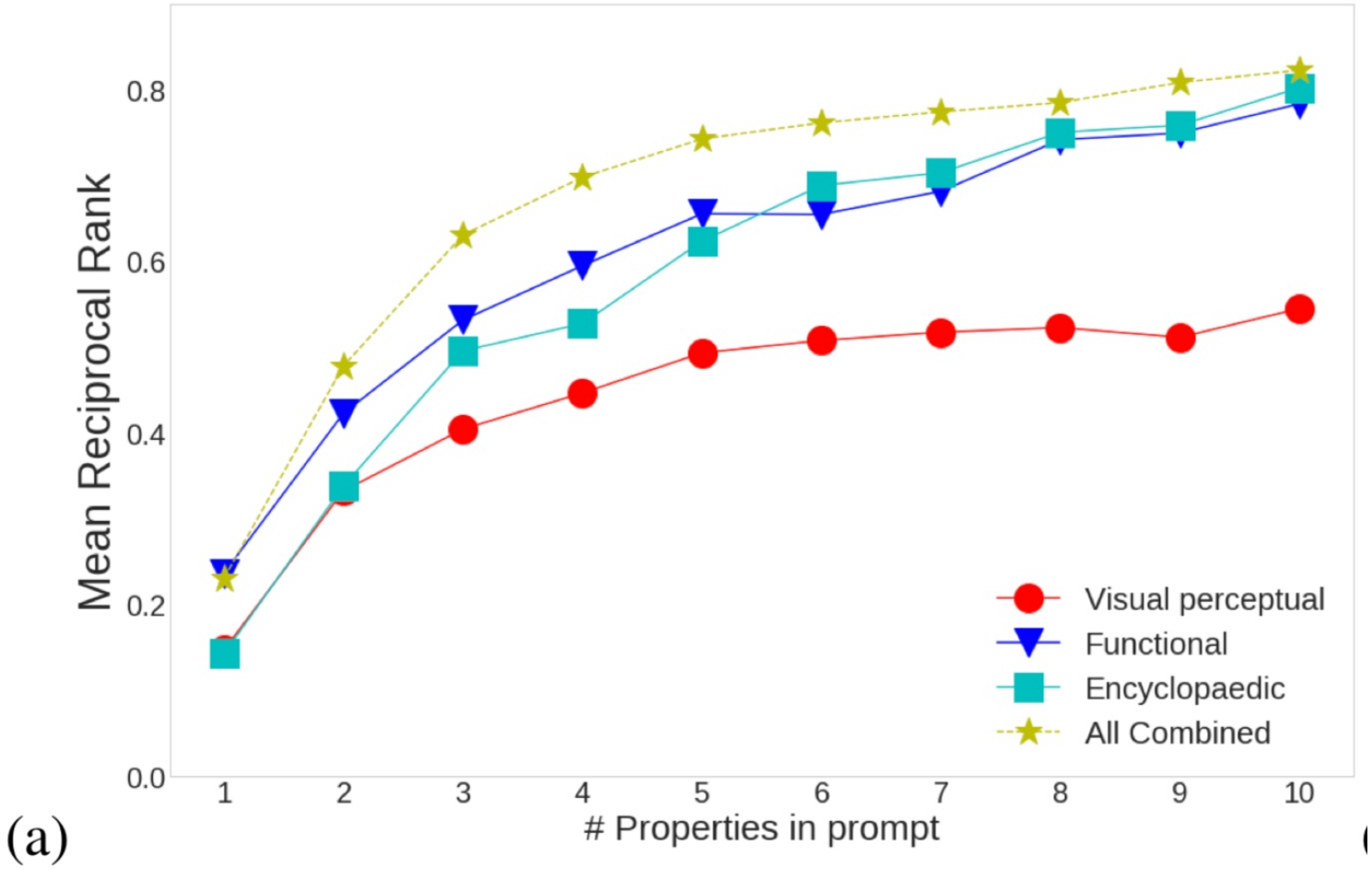
Use: Measure for evaluating any process that produces a list of possible responses to a sample of queries

rank_i is rank position of the *first* relevant document for the *i*-th query,

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}.$$

Distinguish concepts (Weir et al. 2020)

Perceptual cues (bears have fur) < encyclopedic (bears live in forests) [is this surprising?]



Knowledge in Pre-trained LM

- Do pretrained models already have commonsense?
- Knowledge-base completion
- Do pre-trained LMs correctly distinguish concepts associated with a given set of assumed properties?
- Can pre-trained LMs list properties associated with given concepts?

Distinguish concepts (Weir et al. 2020)

Context	Human		RoBERTa-L	
	Response	PF	Response	p_{LM}
(Everyone knows that) a bear has ___ .	fur	27	teeth	.36
	claws	15	claws	.18
	teeth	11	eyes	.05
	cubs	7	ears	.03
	paws	7	horns	.02
(Everyone knows that) a ladder is made of ___ .	metal	25	wood	.33
	wood	20	steel	.08
	plastic	4	metal	.07
	aluminum	2	aluminum	.03
	rope	2	concrete	.03

Low correlation with human elicited properties, but coherent and mostly “verifiable by humans”

Can we trust LMs?

https://demo.allennlp.org/masked-lm

Can we trust LMs?

- LMs generate fictitious facts

Distributionally similar

Barack's Wife Hillary:
Using Knowledge Graphs for Fact-Aware Language Modeling

Robert L. Logan IV* **Nelson F. Liu†§** **Matthew E. Peters§**
Matt Gardner§ **Sameer Singh***

* University of California, Irvine, CA, USA
† University of Washington, Seattle, WA, USA
§ Allen Institute for Artificial Intelligence, Seattle, WA, USA

Negated and Misprimed Probes for Pretrained Language Models:
Birds Can Talk, But Cannot Fly

Syntactically similar

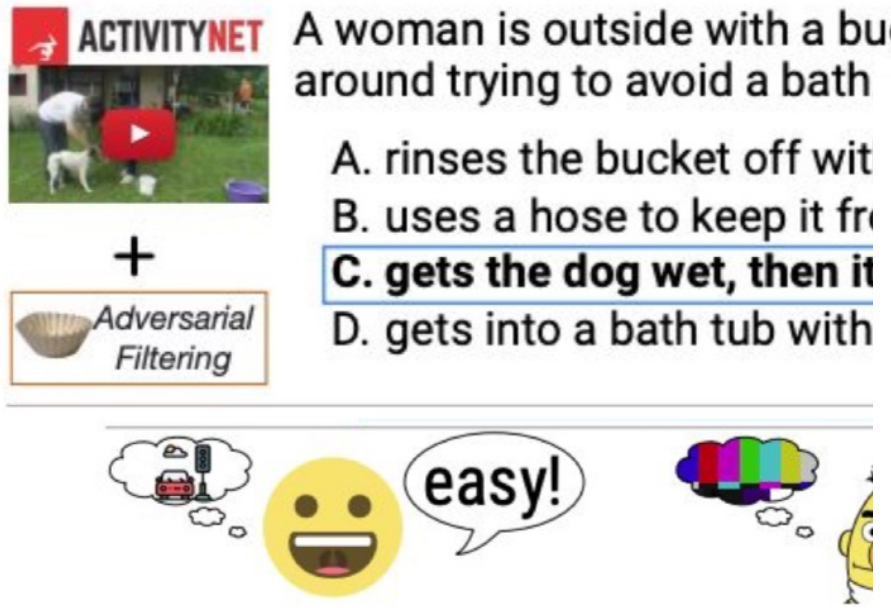
Nora Kassner, Hinrich Schütze
Center for Information and Language Processing (CIS)
LMU Munich, Germany
kassner@cis.lmu.de

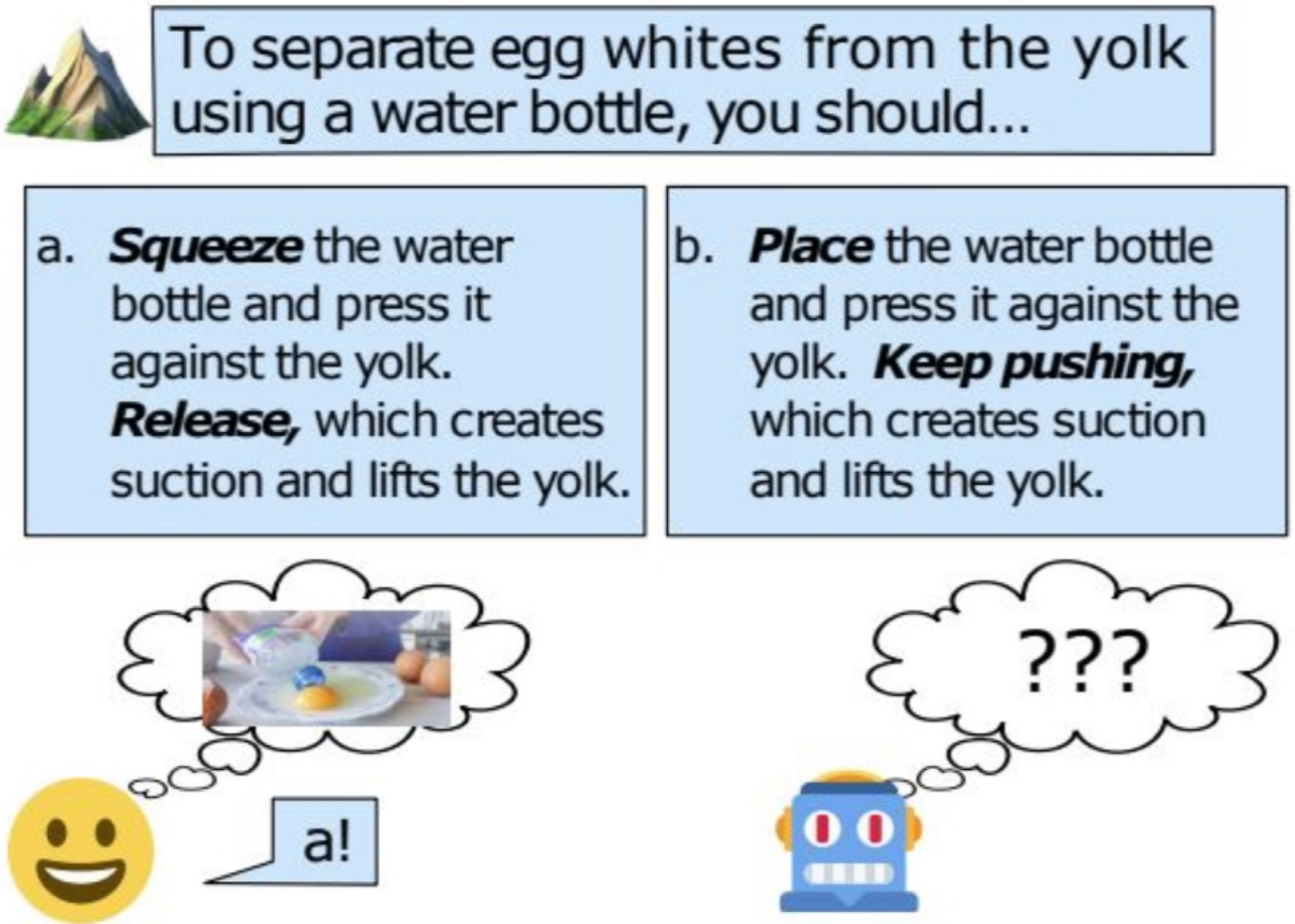
LMs provide a good basis for commonsense models

Performance comes from large pre-training and fine-tuning

- LMs mostly pick up lexical cues
- No model has true commonsense reasoning
  - lack an understanding of some of the most basic physical properties
  - Fails to perform logical reasoning that is critical to commonsense knowledge

Knowledge in LMs isn't enough





Symbolic reasoning

Paradigm in AI

- high-level "symbolic" (human-readable) representations of problems

- NLP: determining whether a *conjunction* of properties is held by an object, and *comparing* the sizes of different objects

Talmor et al. 2020

Probe name	Setup	Example	Human[1]
ALWAYS-NEVER	MC-MLM	*A chicken [MASK] has horns.* **A.** never **B.** rarely **C.** sometimes **D.** often **E.** always	91%
AGE COMPARISON	MC-MLM	*A 21 year old person is [MASK] than me in age, If I am a 35 year old person.* **A.** younger **B.** older	100%
OBJECTS COMPARISON	MC-MLM	*The size of a airplane is [MASK] than the size of a house .* **A.** larger **B.** smaller	100%
ANTONYM NEGATION	MC-MLM	*It was [MASK] hot, it was really cold .* **A.** not **B.** really	90%
PROPERTY CONJUNCTION	MC-QA	*What is usually located at hand and used for writing?* **A.** pen **B.** spoon **C.** computer	92%
TAXONOMY CONJUNCTION	MC-MLM	*A ferry and a floatplane are both a type of [MASK].* **A.** vehicle **B.** airplane **C.** boat	85%
ENCYC. COMPOSITION	MC-QA	*When did the band where Junior Cony played first form?* **A.** 1978 **B.** 1977 **C.** 1980	85%
MULTI-HOP COMPOSITION	MC-MLM	*When comparing a 23, a 38 and a 31 year old, the [MASK] is oldest* **A.** second **B.** first **C.** third	100%

Can LMs do symbolic reasoning

A chicken [MASK] has horns.

A. never B. rarely C. sometimes D. often E. always

Talmor et al. (2020): oLMpics for BERT and RoBERTa on a set of symbolic reasoning tasks

Neither perform well

Reporting bias: LMs are trained on texts describing things that ***do*** happen

Symbolic reasoning

	RoBERTa Large	BERT WWM	BERT Large	RoBERTa Base	BERT Base
ALWAYS-NEVER					
AGE COMPARISON	✓	✓		✓✗	
OBJECTS COMPAR.	✓	✓✗			
ANTONYM NEG.	✓		✓✗	✓✗	
PROPERTY CONJ.	✓✗	✓✗			
TAXONOMY CONJ.	✓✗	✓✗		✓✗	
ENCYC. COMP.					
MULTI-HOP COMP.					

Talmor et al. 2020

Some commonsense knowledge but very far from being complete

Table 12: The oLMpic games medals', summarizing per-task success. ✓ indicate the LM has achieved high accuracy considering controls and baselines, ✓✗ indicates partial success.

How do we measure commonsense reasoning

Benchmark tasks

How do you know that a model is doing commonsense reasoning?

Unsupervised:

- Observe behavior,
- Probe representations,
- etc.

Benchmarks:
knowledge-specific tests
(w/ or w/o training data)

QA format: easy to evaluate
(e.g., accuracy)

Step 1: Determine type of reasoning

Abductive reasoning

Visual commonsense reasoning

https://leaderboard.allenai.org/

Social
IQa

Reasoning about Social Situations

Alex spilt food all over the floor and it made a huge mess.

What will Alex want to do next?

run around in the mess

less likely

mop up the mess

more likely

Step 2: Choosing a benchmark size

	Small scale	**Large scale**
Creation	Expert-curated	Crowdsourced/automatic
Coverage	Limited coverage	Large coverage
Training	Dev/test only	Training/dev/test
Budget	Expert time costs	Crowdsourcing costs

Winograd Schema Challenge (WSC),
Choice of Plausible Alternatives (COPA)

Small commonsense benchmarks

Winograd Schema Challenge (WSC)
273 examples

Choice of Plausible Alternatives (COPA)
500 dev, 500 test

The city councilmen refused the demonstrators a permit because *they* **advocated** violence. Who is "*they*"?

(a) The city councilmen
(b) The demonstrators

The city councilmen refused the demonstrators a permit because *they* **feared** violence. Who is "*they*"?

(a) The city councilmen
(b) The demonstrators

Small commonsense benchmarks

Winograd Schema Challenge (WSC)
273 examples

Choice of Plausible Alternatives (COPA)
500 dev, 500 test

I hung up the phone.
What was the **cause** of this?

(a) The caller said goodbye to me.
(b) The caller identified himself to me.

The toddler became cranky.
What happened as a **result**?

(a) Her mother put her down for a nap.
(b) Her mother fixed her hair into pigtails.

Step 2: Choosing a QA benchmark size

	Small scale	**Large scale**
Creation	Expert-curated	Crowdsourced/automatic
Coverage	Limited coverage	Large coverage
Training	Dev/test only	Training/dev/test
Budget	Expert time costs	Crowdsourcing costs

Challenge: do to collect positive/negative answers?

Challenge of collecting unlikely answers

Goal: negative answers have to be ***plausible but unlikely***

- Automatic matching?
  - Random negative sampling won't work, too topically different
  - "smart" negative sampling isn't effective either
- Need better solution… maybe we can ask crowd workers?

Collecting answers from crowdworkers

Social IQa

Context and Question

Alex spilt food all over the floor and it made a huge mess.

WHAT HAPPENS NEXT

What will Alex want to do next?

amazon mturk Requester

Free Text Response

Handwritten ✓ and ✗ Answers

✓ mop up
✓ give up and order take out
✗ leave the mess
✗ run around in the mess

Problem: handwritten unlikely answers are **too easy to detect**

Problem: annotation artifacts

- Models can exploit artifacts in handwritten incorrect answers
  - Exaggerations, off-topic, overly emotional, etc.
  - See Schwartz et al. 2017, Gururangan et al. 2018, Zellers et al. 2018, etc.
- Seemingly "super-human" performance by large pretrained LMs (BERT, GPT, etc.)

Benchmark creation important to avoid overstating performance ("super-human machine")

Commonsense benchmarks

Social commonsense

Naïve Psychology

ROC story

Social IQa

WSC

COPA

Abductive NLI

VCR

WinoGrande

CommonsenseQA

Physical commonsense

Physical IQa

HellaSwag

SWAG

JHU Ordinal Commonsense

MCTaco

Temporal commonsense

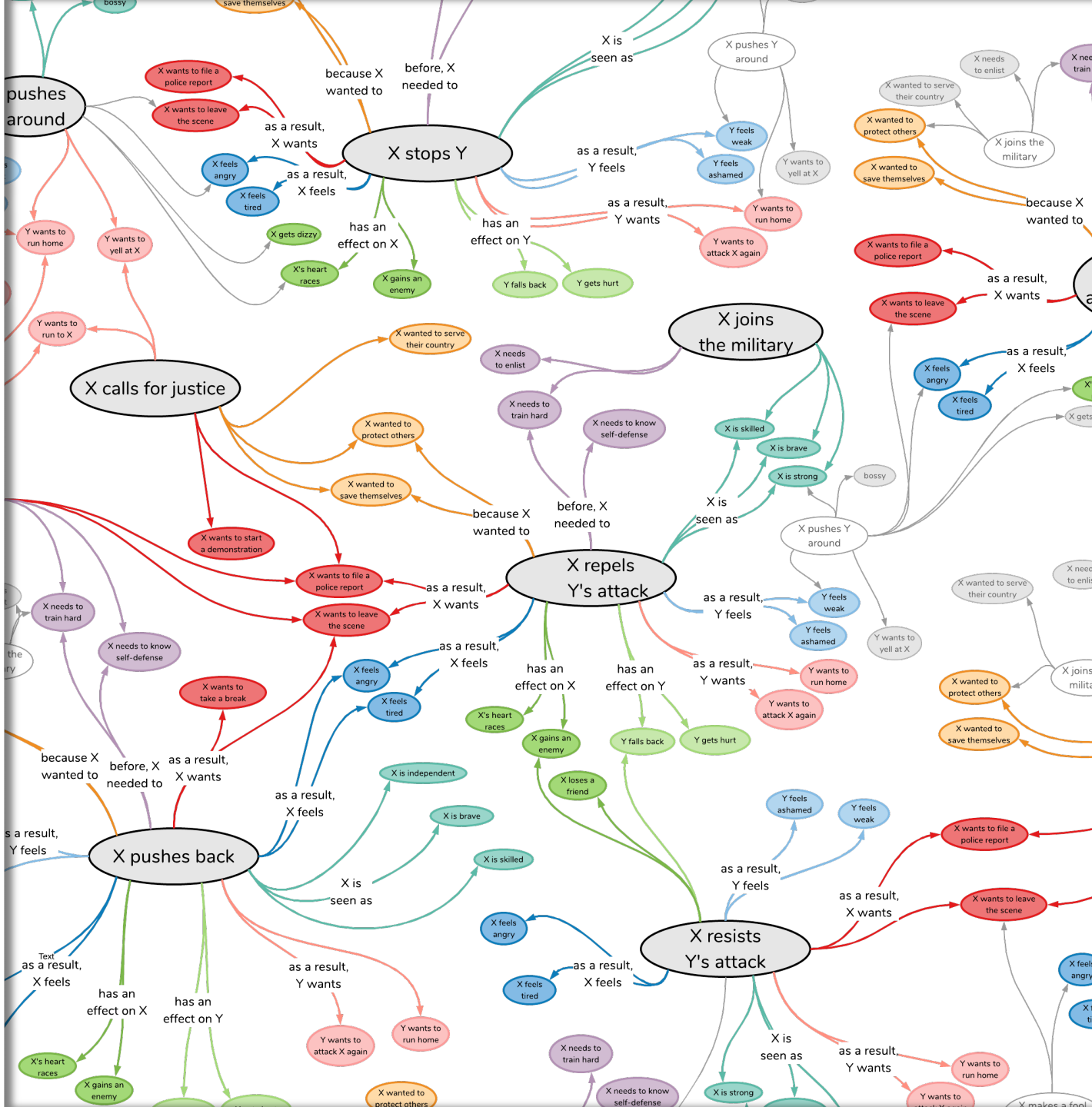
ReCORD

CosmosQA

MultiRC

Commonsense reading comprehension

Commonsense resources



Grandma’s glasses

Tom’s grandma was reading a new book, when she dropped her glasses.

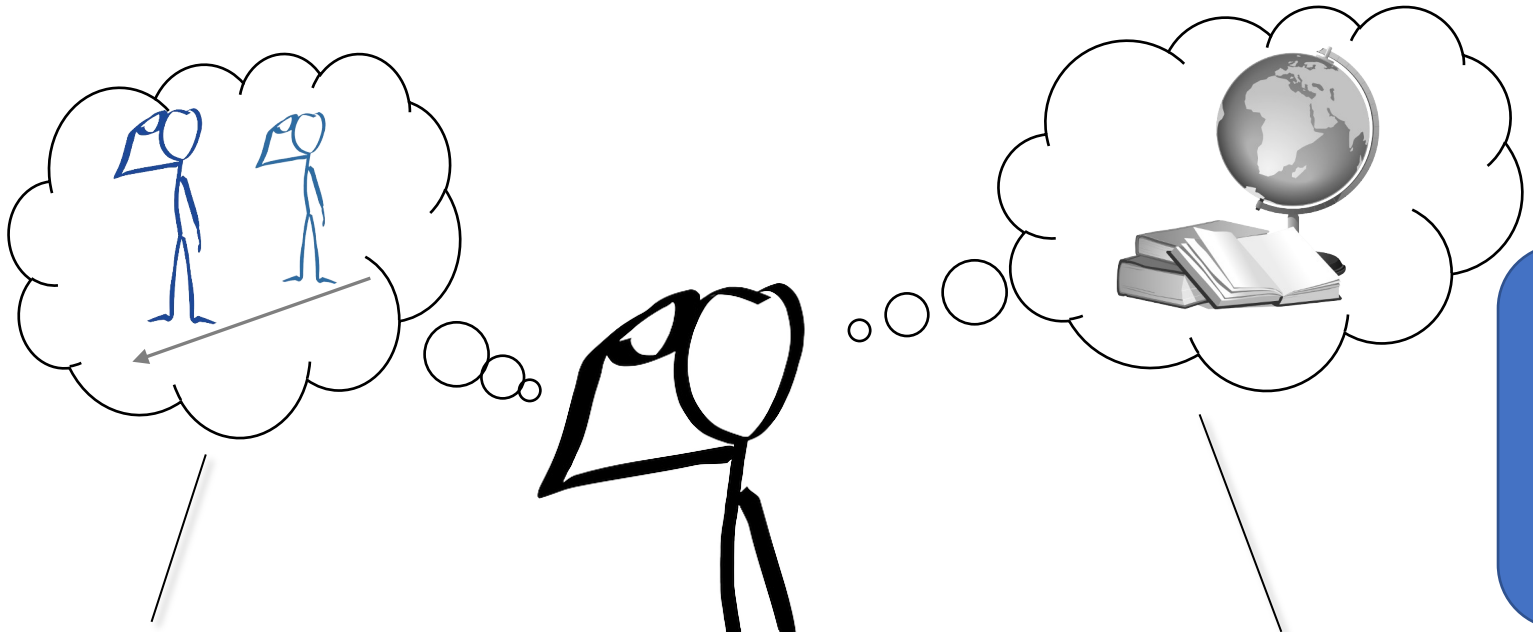
She couldn’t pick them up, so she called Tom for help.

Tom rushed to help her look for them, they heard a loud crack.

They realized that Tom broke her glasses by stepping on them.

Promptly, his grandma yelled at Tom to go get her a new pair.

Humans reason about the world with **mental models** [Graesser, 1994]



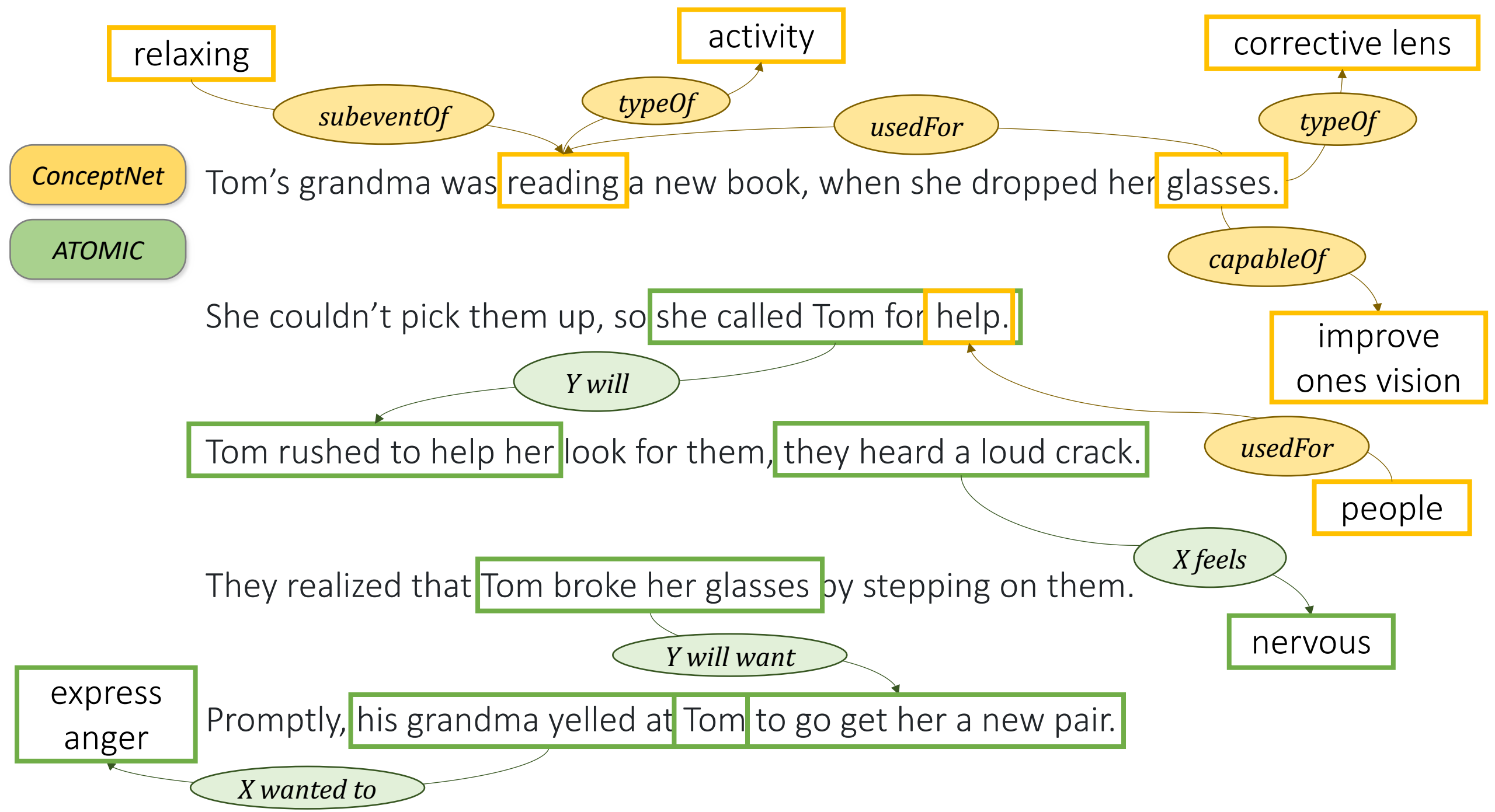
Personal experiences

[Conway et al., 2000]

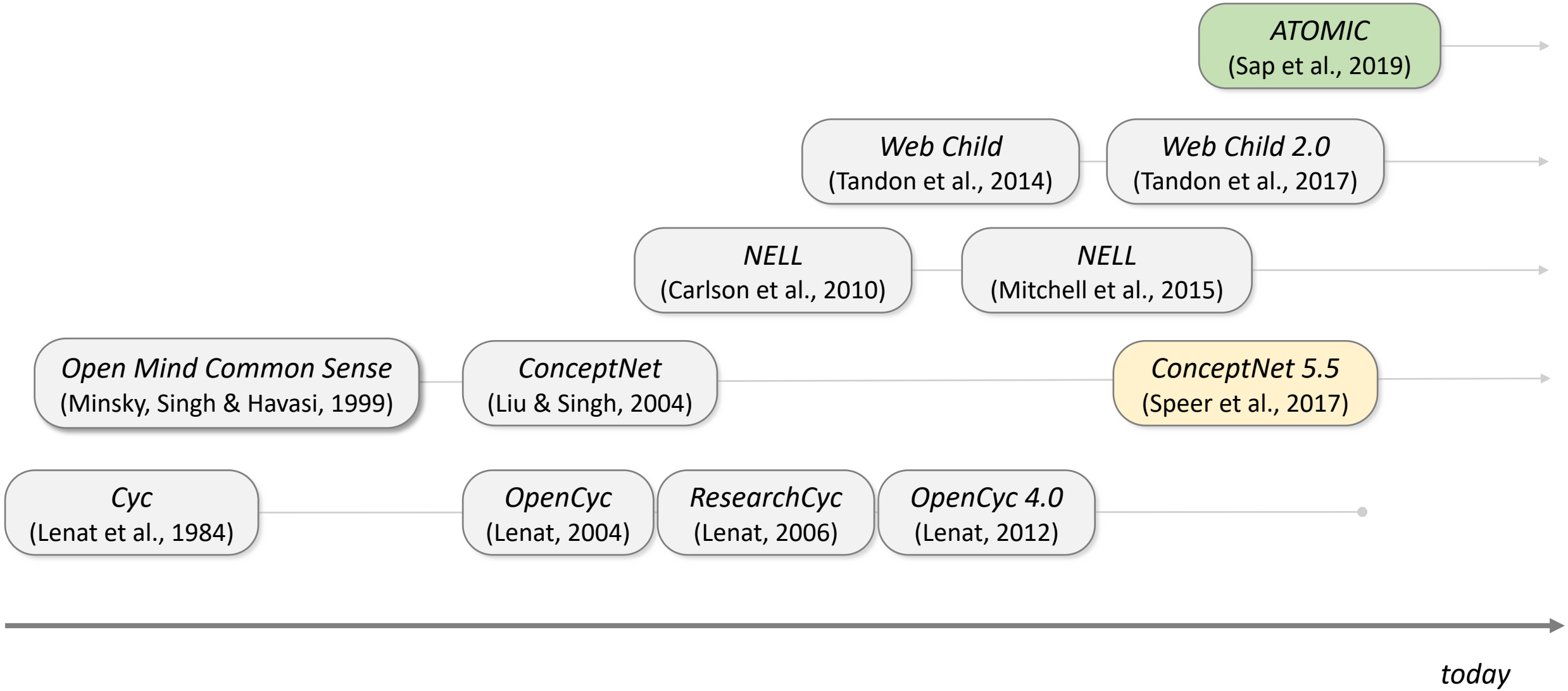
World knowledge and commonsense

[Kintsch, 1988]

Commonsense resources aim to be a bank of knowledge for machines to be able to reason about the world in tasks


relaxing
activity
corrective lens
subeventOf
typeOf
usedFor
typeOf
ConceptNet
Tom's grandma was reading a new book, when she dropped her glasses.
ATOMIC
capableOf
She couldn't pick them up, so she called Tom for help.
improve ones vision
Y will
Tom rushed to help her look for them, they heard a loud crack.
usedFor
people
X feels
They realized that Tom broke her glasses by stepping on them.
nervous
Y will want
express anger
Promptly, his grandma yelled at Tom to go get her a new pair.
X wanted to

Overview of existing resources



How do you create a commonsense resource?

Desiderata for a good commonsense resource

Coverage

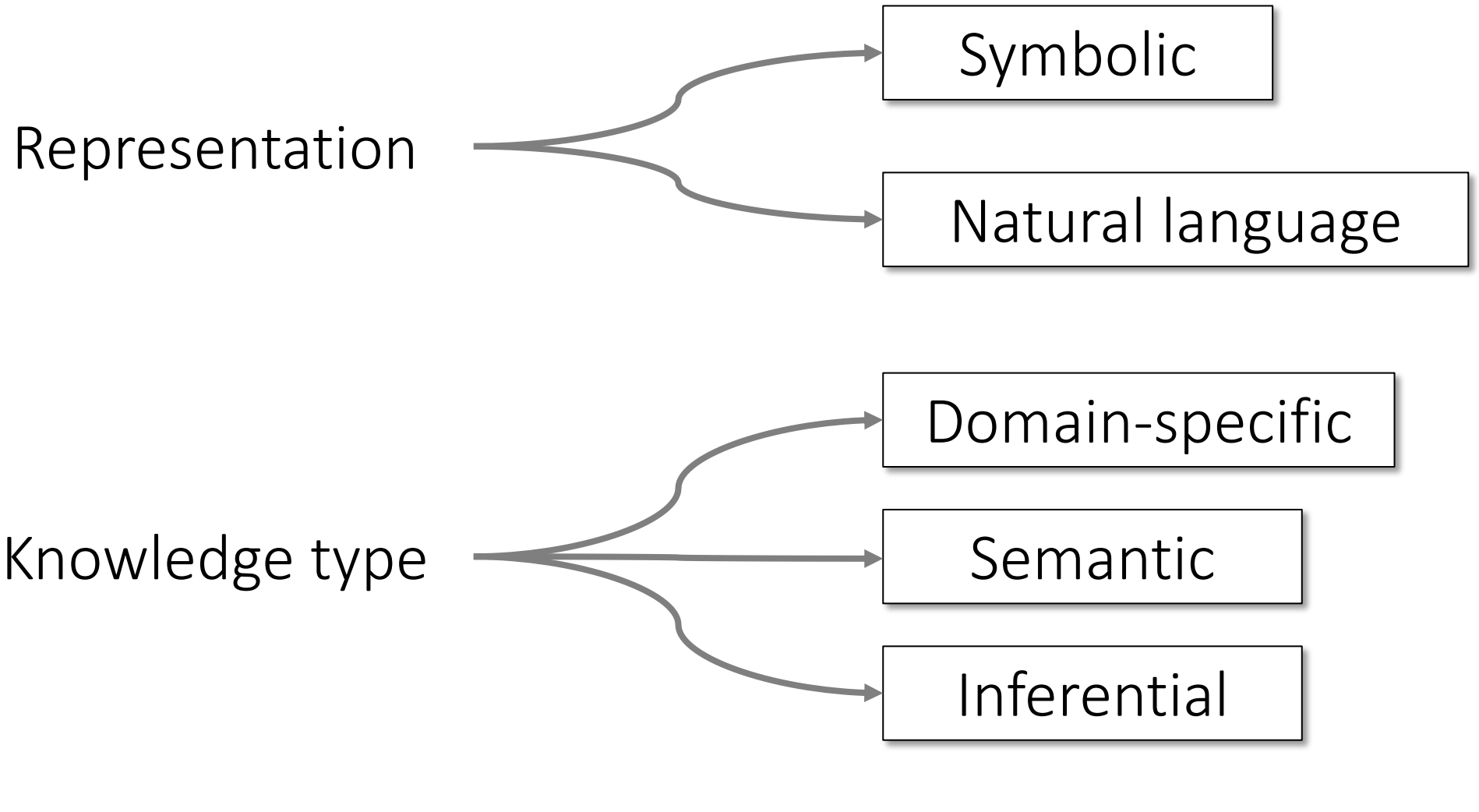
- Large scale
- Diverse knowledge types

Useful

- High quality knowledge
- Usable in downstream tasks

Multiple resources tackle different knowledge types

Creating a commonsense resource



CONCEPTNET:

semantic knowledge in *natural language* form

http://conceptnet.io/

en reading

An English term in ConceptNet 5.8

Related terms

- en book →
- en books →
- en book →

Effects of reading

- en learning →
- en ideas →
- en a headache →

reading is a type of...

- en an activity →
- en a good way to learn →
- en one way of learning →
- en one way to learn →

reading is a subevent of...

- en you learn →
- en turning a page →
- en learning →

Subevents of reading

- en relaxing →
- en study →
- en studying for a subject →

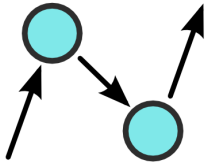
Things used for reading

- en article →
- en a library →
- en literature →
- en a paper page →

Types of reading

- en browse (n, communication) →
- en bumf (n, communication) →
- en clock time (n, time) →
- en miles per hour (n, time) →

What is ConceptNet?



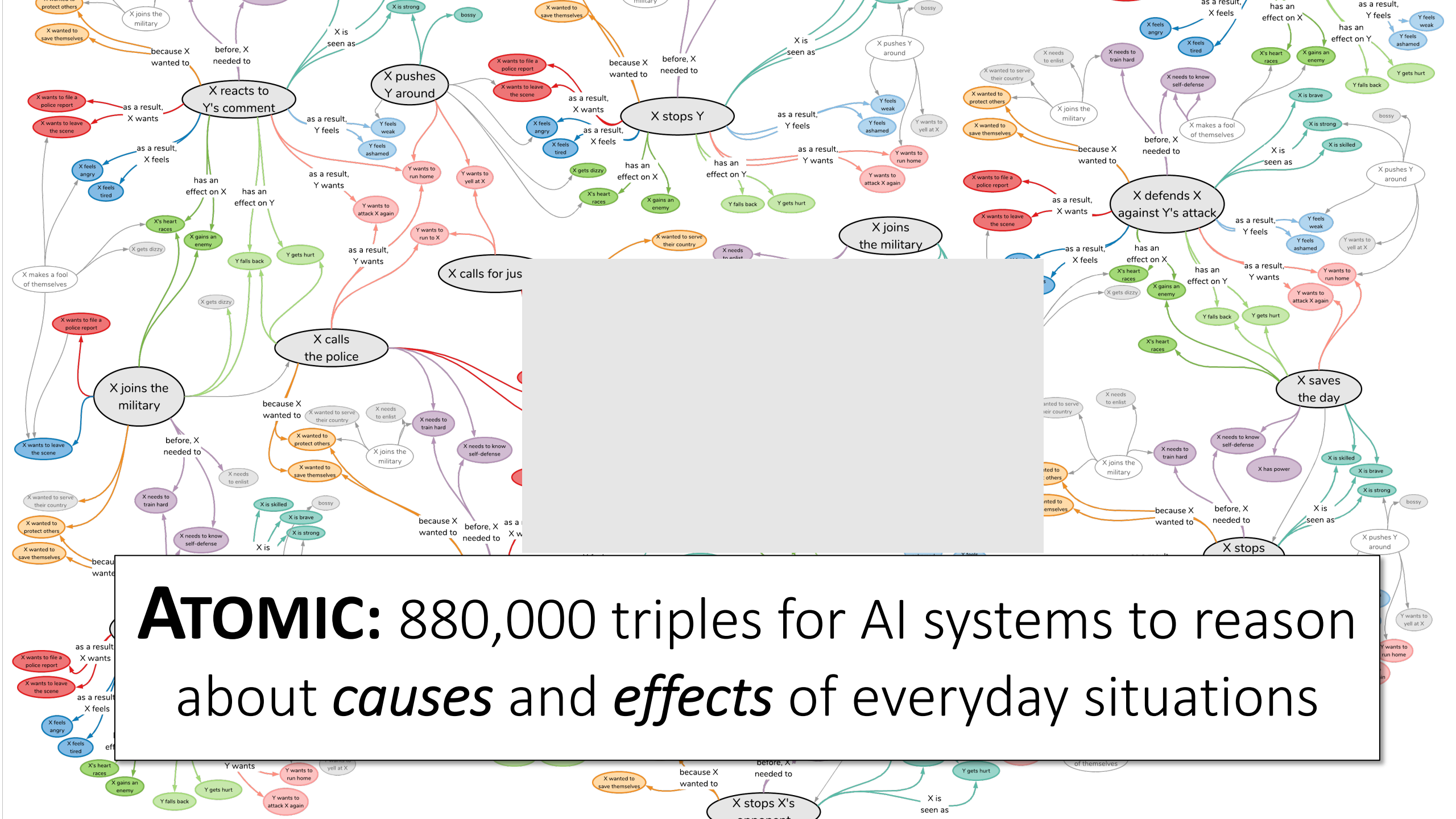
- General commonsense knowledge
- 21 million edges and over 8 million nodes (as of 2017)
  - Over 85 languages
  - In English: over 1.5 million nodes
- Knowledge covered:
  - Open Mind Commonsense assertions
  - Wikipedia/Wiktionary semantic knowledge
  - WordNet, Cyc ontological knowledge

http://conceptnet.io/

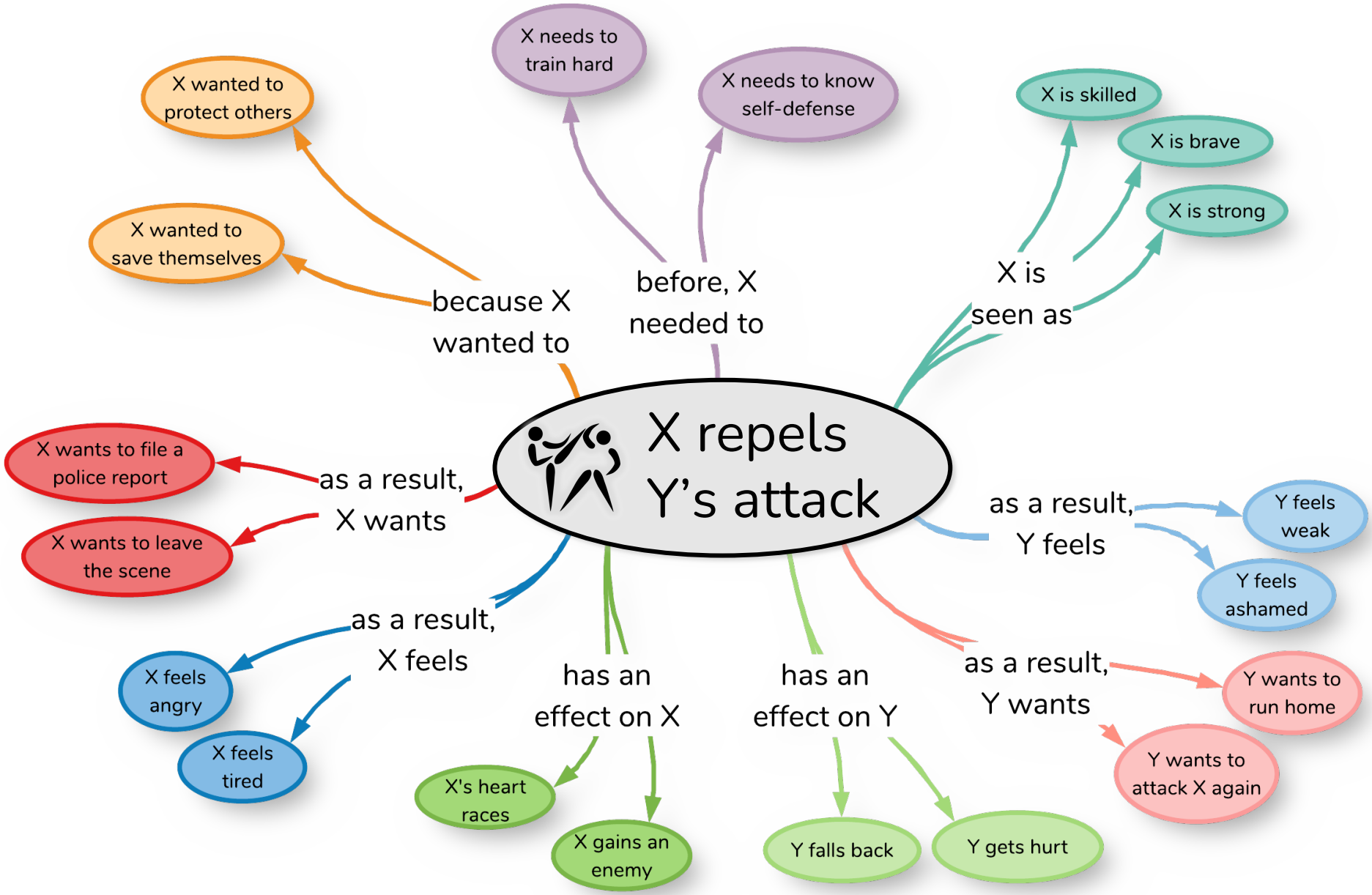
ATOMIC:

inferential knowledge in *natural language* form

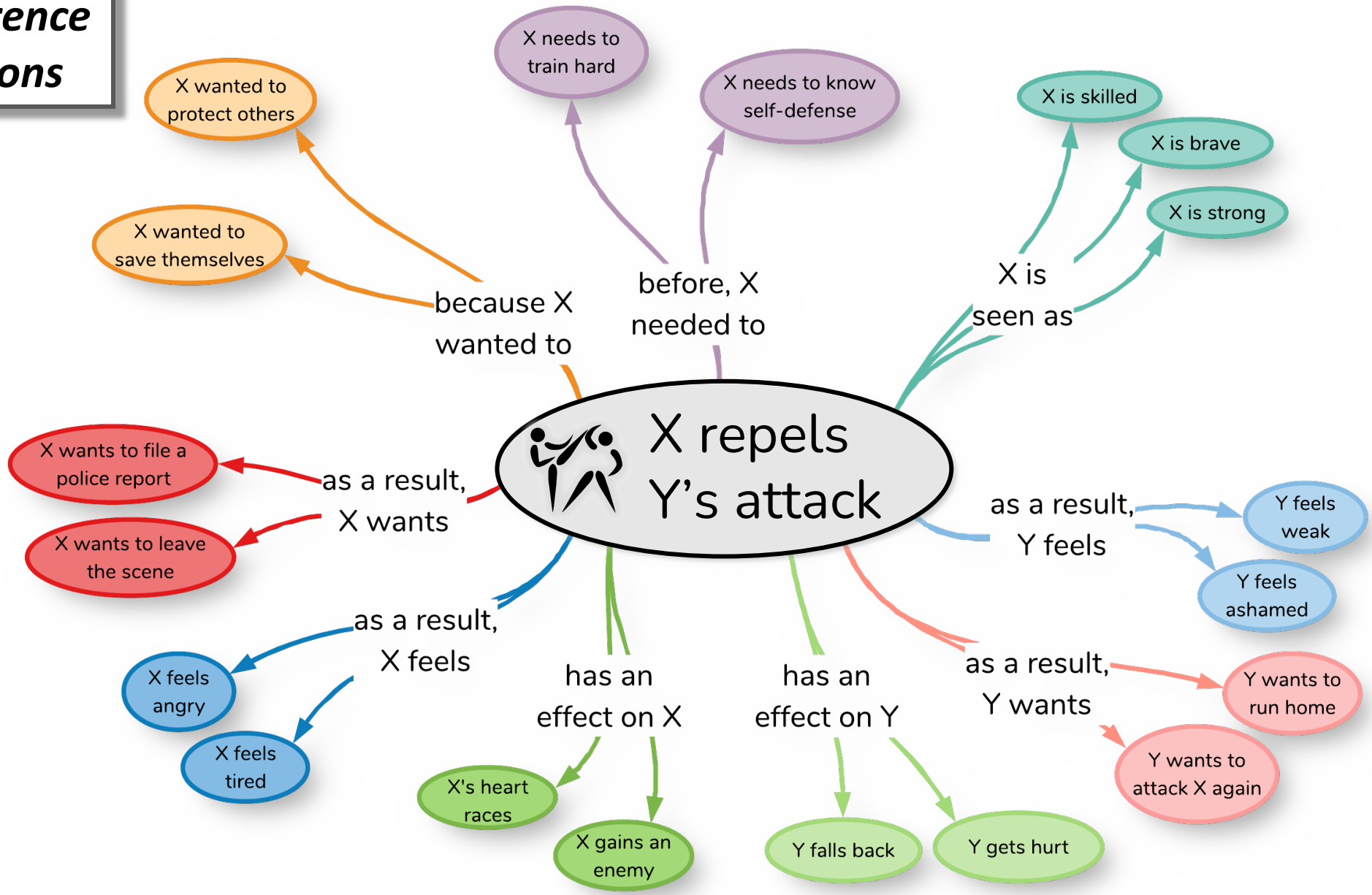
https://mosaickg.apps.allenai.org/kg_atomic

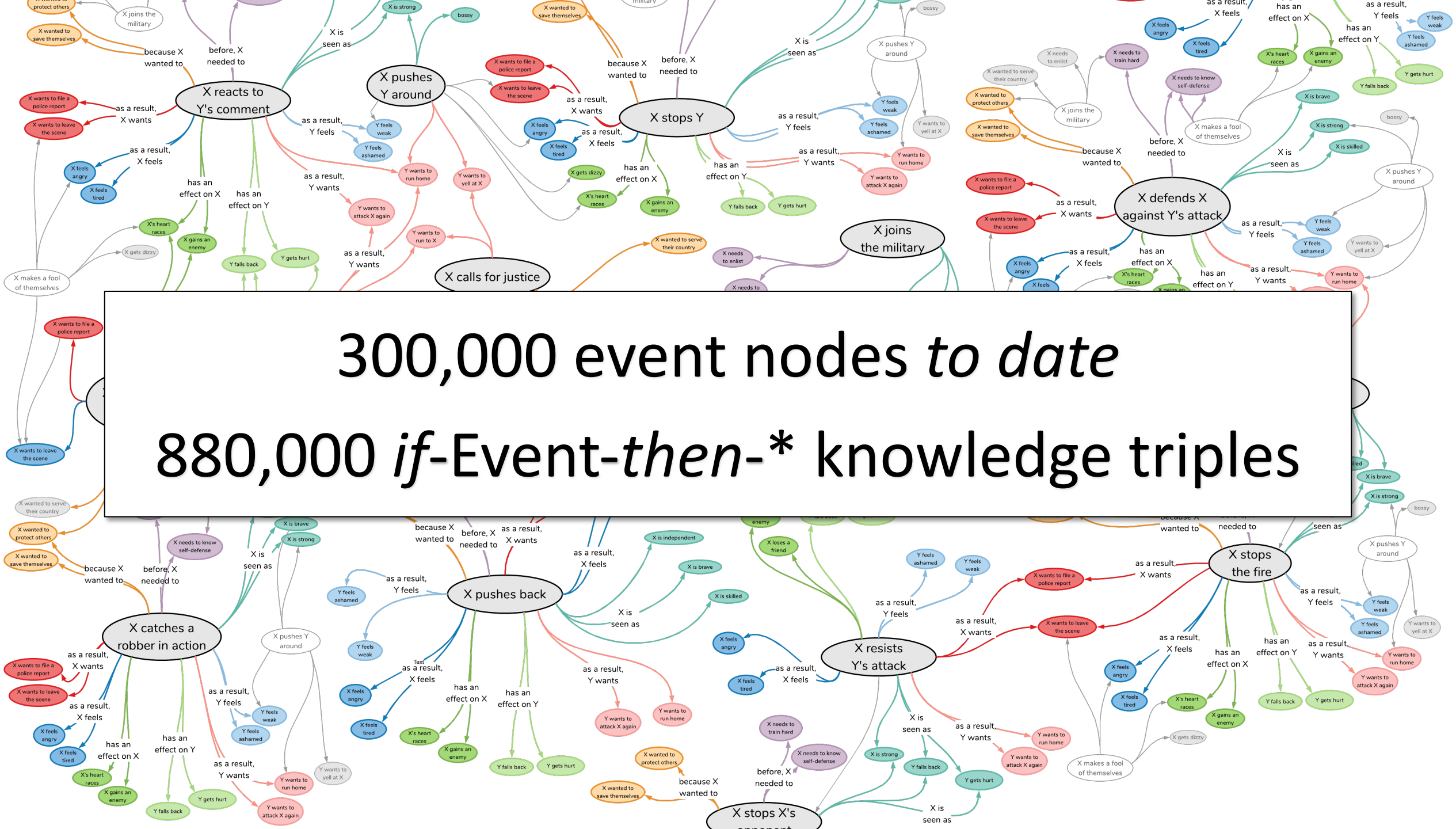


ATOMIC: 880,000 triples for AI systems to reason about ***causes*** and ***effects*** of everyday situations

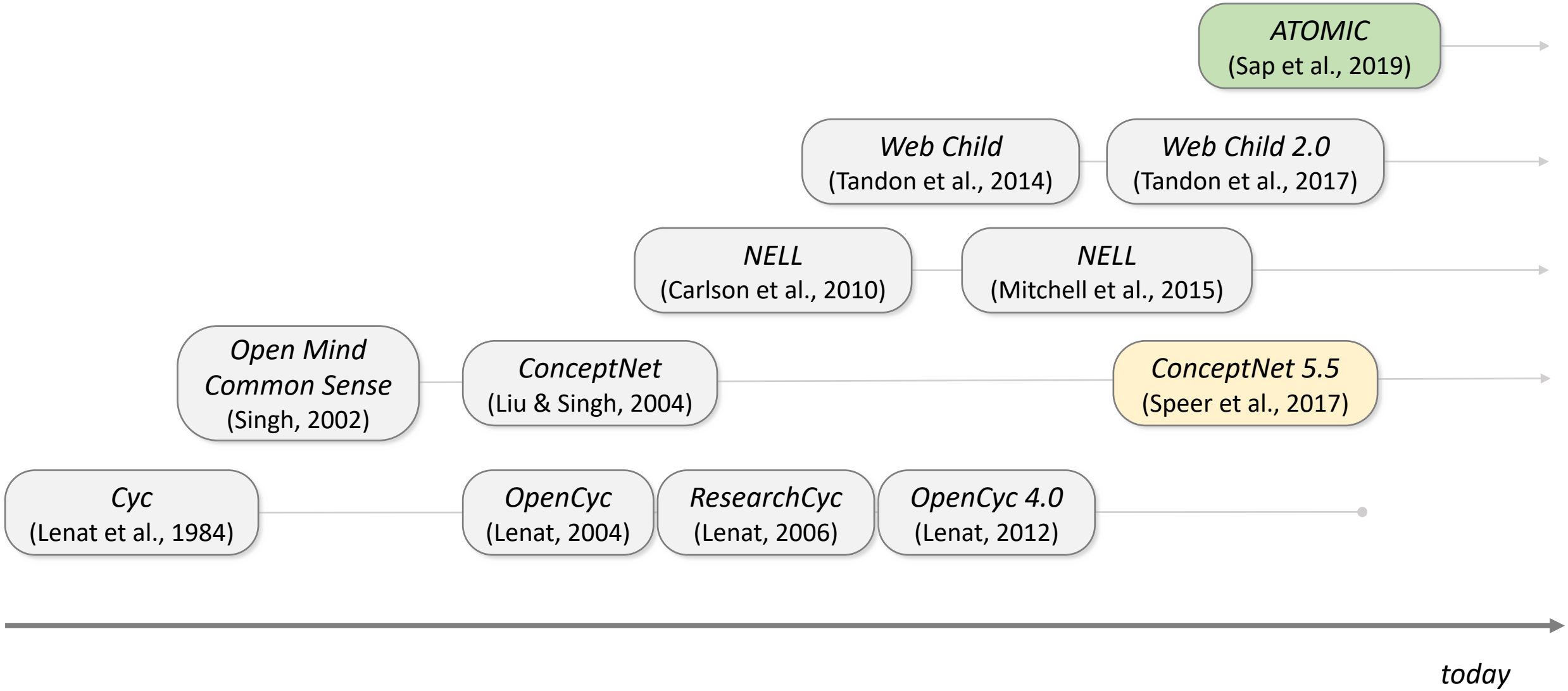

X repels Y's attack
because X wanted to
X wanted to protect others
X wanted to save themselves
before, X needed to
X needs to train hard
X needs to know self-defense
X is seen as
X is skilled
X is brave
X is strong
as a result, X wants
X wants to file a police report
X wants to leave the scene
as a result, X feels
X feels angry
X feels tired
has an effect on X
X's heart races
X gains an enemy
has an effect on Y
Y falls back
Y gets hurt
as a result, Y wants
Y wants to run home
Y wants to attack X again
as a result, Y feels
Y feels weak
Y feels ashamed

nine inference dimensions




300,000 event nodes *to date*
880,000 *if*-Event-*then*-* knowledge triples

Overview of existing resources



Existing knowledge bases

ATOMIC
(Sap et al., 2019)

NELL
(Mitchell et al., 2015)

ConceptNet 5.5
(Speer et al., 2017)

OpenCyc 4.0
(Lenat, 2012)

Existing knowledge bases

Represented in **symbolic logic**
(e.g., LISP-style logic)

Represented in **natural language**
(how humans *talk* and *think*)

NELL
(Mitchell et al., 2015)

OpenCyc 4.0
(Lenat, 2012)

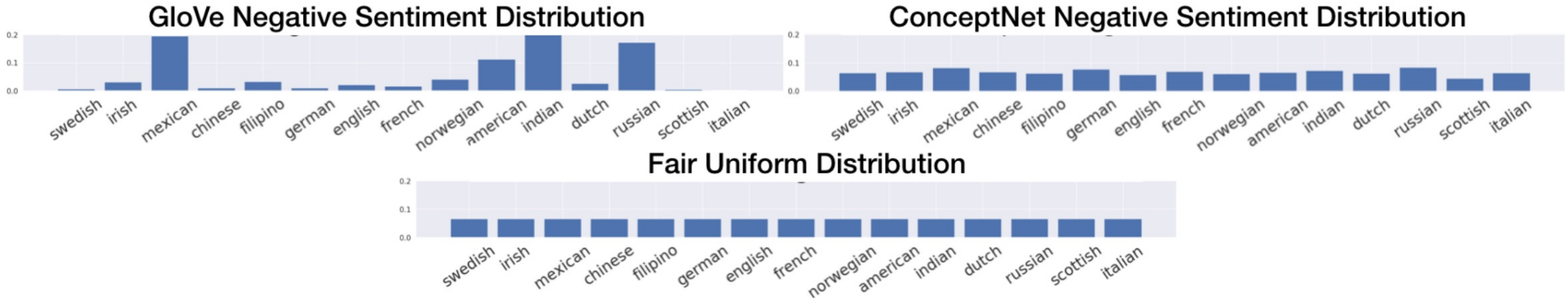
ConceptNet 5.5
(Speer et al., 2017)

ATOMIC
(Sap et al., 2019)

```
(#$implies
  (#$and
    (#$isa ?OBJ ?SUBSET)
    (#$genls ?SUBSET ?SUPERSET))
  (#$isa ?OBJ ?SUPERSET))
```

Knowledge bases and mitigating biases

- Different data collection methods suffer from social biases differently
- ConceptNet word embeddings have less demographic biases than GloVe embeddings [Sweeney & Najafian, 2019]



Knowledge bases and mitigating biases

PersonX clutches a gun → because X wanted to
- to be safe
- to protect himself
- to protect themselves
- to defend themselves
- to defend himself

ATOMIC (Sap et al., 2019)

Jaquain clutches a gun → because X wanted to
- to kill someone
- none
- to protect himself
- to be safe
- to protect themselves

Karen clutches a gun → because X wanted to
- to be safe
- to protect himself
- to shoot
- to get the gun
- none

COMET (Bosselut et al., 2019): ATOMIC + OpenAI GPT

Some commonsense cannot be extracted

Text is subject to **reporting bias**
(Gordon & Van Durme, 2013)

- Idioms & figurative usage

  "Black sheep problem"

- Noteworthy events

  Murdering 4x more common than exhaling

Commonsense is not often written

-> Grice's maxim of quantity

found when extracting commonsense knowledge on four large corpora using Knext (Gordon & Van Durme, 2013)

How do we Incorporate Commonsense into Downstream Models?

Katrina had the financial means to afford a new car while Monica did not, since ____ had a high paying job.

WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale.
Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi.
AAAI 2020

[CLS] Katrina had the financial means to afford a new car while Monica did not, since [SEP] Katrina had a high paying job.

0.51
0.49

WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale.
Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi.
AAAI 2020

Fine-tuning is important

Sentence:

Katrina had the financial means to afford a new car while Monica did not, since [MASK] had a high paying job.

Predictions:

11.8% ↵

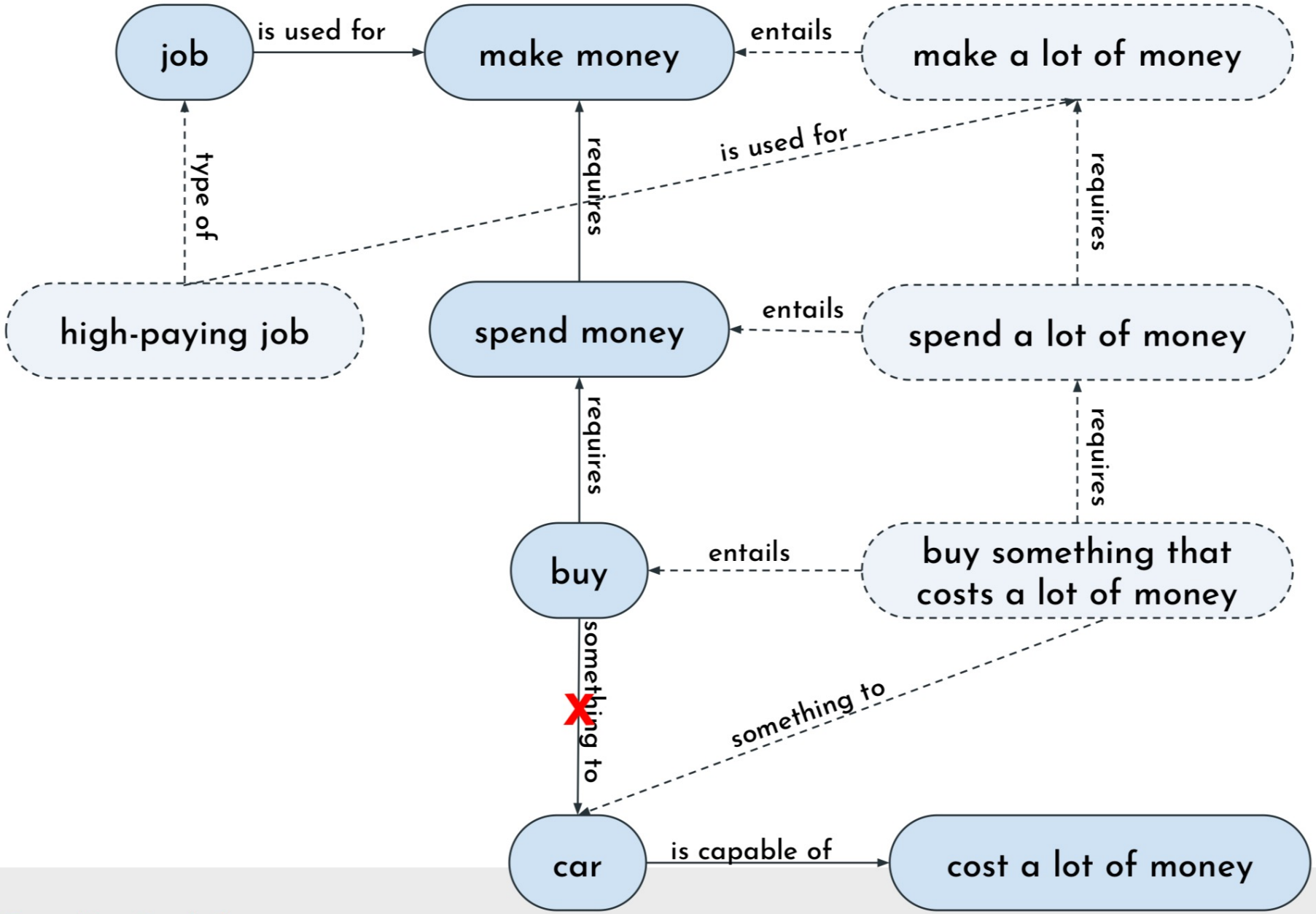
8.8% **She**

6.3% **I**

6.2% **So**

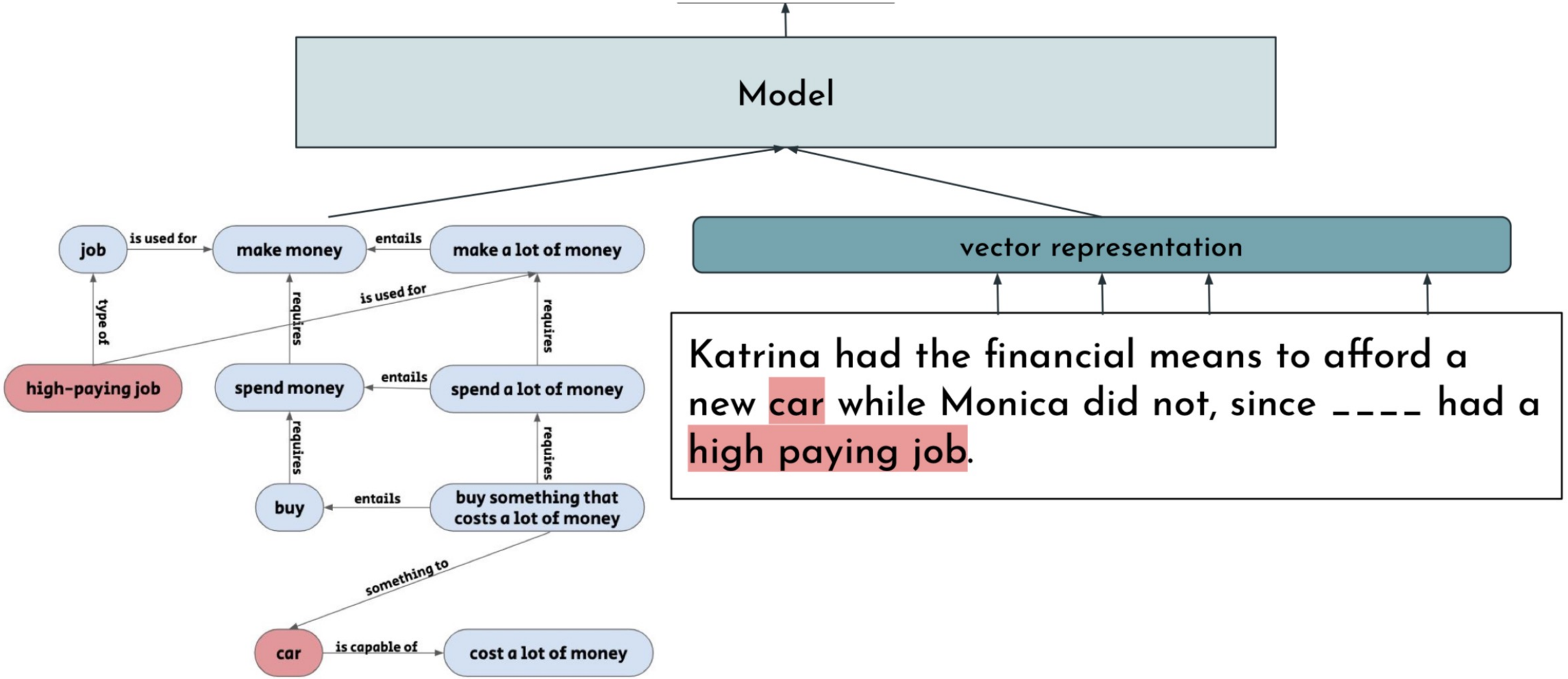
5.2% **Monica**

← **Undo**



http://conceptnet5.media.mit.edu/

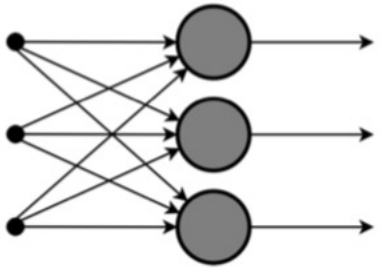
Static KB integration



Recipe

Task

Story ending,
Machine Comprehension
Social common sense
NLI

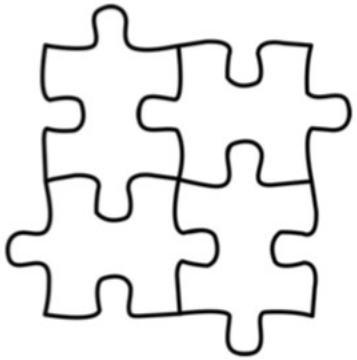


Neural Component

Pre/post pre-trained
language models

Knowledge Source

Knowledge bases,
extracted from text,
hand-crafted rules



Combination Method

Attention, pruning, word
embeddings, multi-task
learning

Tasks

ProPara

Paragraph (seq. of steps):		Participants: water	light	CO2	mixture	sugar
	state0	soil	sun	?	-	-
Roots absorb water from soil						
	state1	roots	sun	?	-	-
The water flows to the leaf.						
	state2	leaf	sun	?	-	-
Light from the sun and CO2 enter the leaf.						
	state3	leaf	leaf	leaf	-	-
The light, water, and CO2 combine into a mixture.						
	state4	-	-	-	leaf	-
Mixture forms sugar.						
	state5	-	-	-	-	leaf

Time

MCScript

T I wanted to plant a tree. I went to the home and garden store and picked a nice oak. Afterwards, I planted it in my garden.

Q1 What was used to dig the hole?
a. a shovel b. his bare hands

Q2 When did he plant the tree?
a. after watering it b. after taking it home

NarrativeQA

Question: How is Oscar related to Dana?
Answer: her son
Snippet: [...] She continues digging in her purse while Frank leans over the buggy and makes funny faces at the baby, OSCAR, a very cute nine-month old boy. [...]

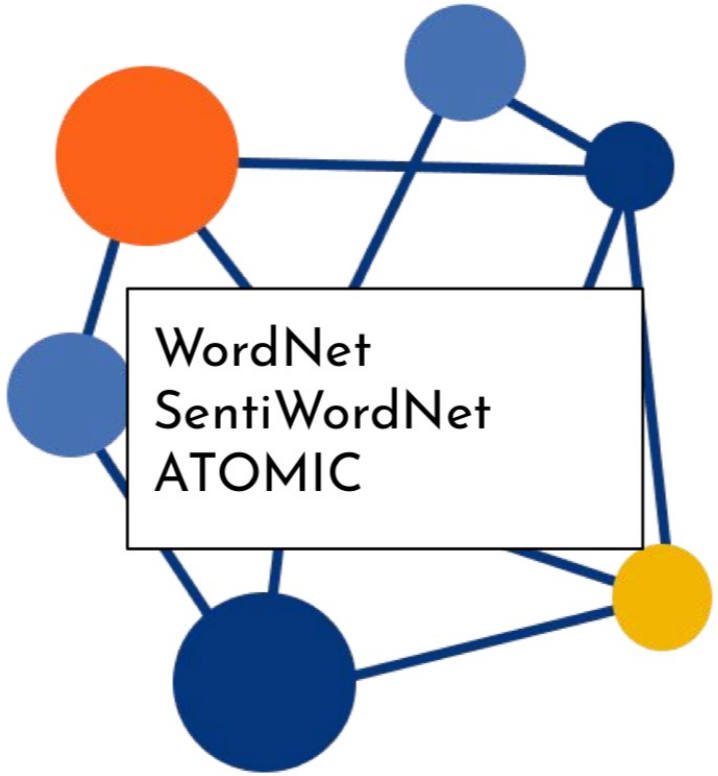
Tasks

Agatha had always wanted pet birds.
So one day she purchased two pet finches.
Soon she couldn't stand their constant noise.
And even worse was their constant mess.

Agatha decided to buy two more. (Wrong)
Agatha decided to return them. (Right)

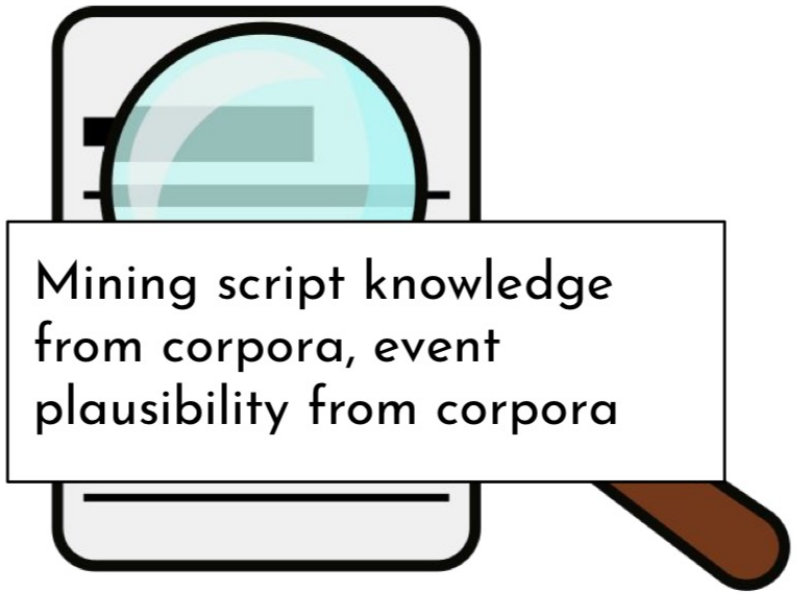
A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. *Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen.* NAACL 2016.

Knowledge Sources



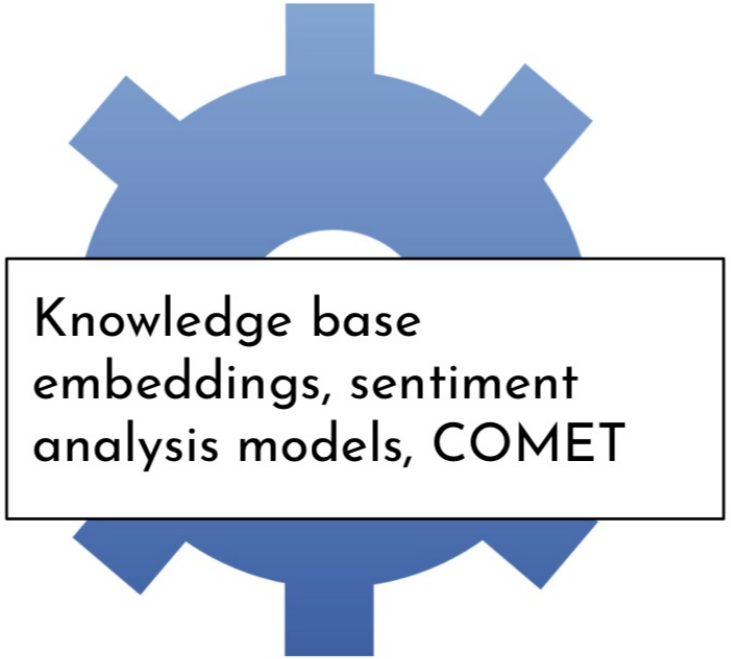
WordNet
SentiWordNet
ATOMIC

Knowledge Bases



Mining script knowledge from corpora, event plausibility from corpora

Mining from Text



Knowledge base embeddings, sentiment analysis models, COMET

Tools

Neural component

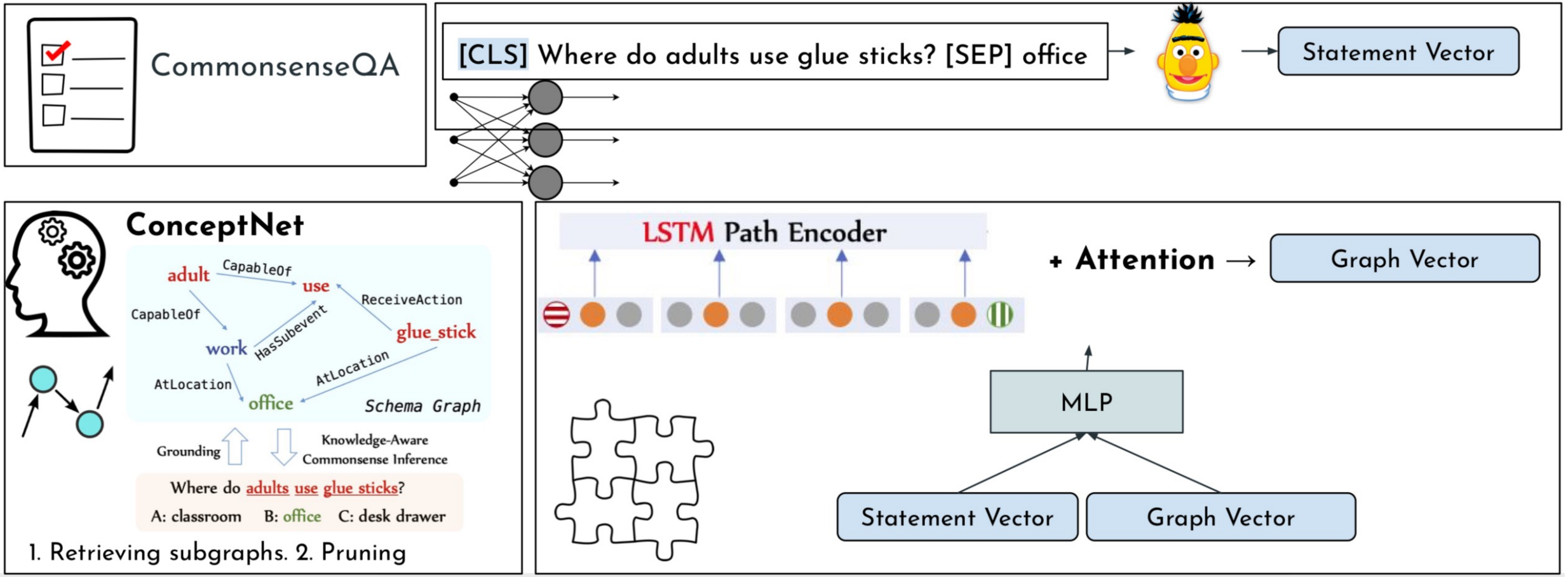
[CLS] Katrina had the financial means to afford a new car while Monica did not, since [SEP] Katrina had a high paying job.

0.51
0.49

Combination Method

1. Incorporate into scoring function
2. Symbolic → vector representation
   - (+attention)
3. Multi-task learning

Combination Method



KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. *Bill Yuchen Lin, Xinyue Chen, Jamin Chen, Xiang Ren.* EMNLP 2019.

Limitation of KB approaches

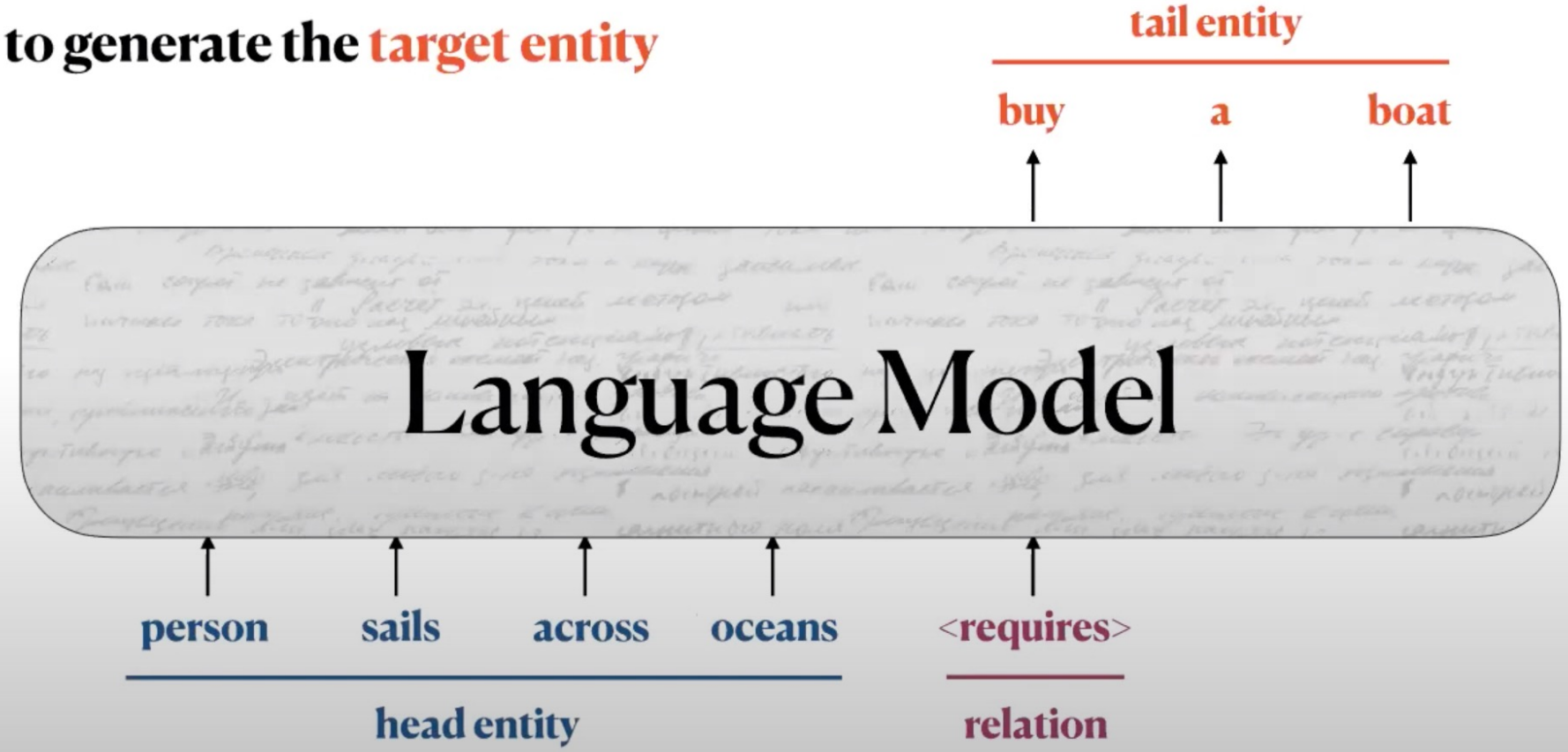
- Situations rarely as in KBs
  - KB only a snapshot of vast commonsense knowledge

- Solutions
  - Learn from KBs, and induce new relationships
  - Scale-up using language resources

Going beyond KBs

$$\mathcal{L} = -\sum \log P(\textbf{target words} \mid \textbf{seed words, relation})$$

Given a seed entity and a relation, learn to generate the target entity



(Bosselut et al., 2019)

Visual Commonsense Knowledge Graphs

Sink in the water.

Try to help

Save himself from

Wait for help to arrive.

Notice water washing in.

Swim towards the statute.

Swim to safety.

After Person1

Because Person2

Sense his own death.

Get to the top of the

Because Person1

Before Person2

Be washed away.

Realize the ship is

Before Person1

After Person2

Start moving against the

Get caught in a rush of water.

Gasp for

Scream for help.

Park et al., 2020

Static vs. Dynamic

Kai knew that things were getting out of control and managed to keep his temper in check

Link to static Knowledge Graph

bad links

X keeps ___ under control

X keeps X's temper

X sweats

X wants to show

X avoids a fight

X keeps X's ___ in check

context-free knowledge

Generate dynamic graph with COMET

Kai stays calm

Kai is viewed as cautious

no linking

Kai wants to avoid trouble

Kai intends

contextual knowledge

Summary

- What do LMs know (very little)?
- How do we measure commonsense ability (benchmarks)?
- What are commonsense resources for machines (representation)?
- How to integrate commonsense into machines?