Logistics

- Project progress presentation 3/30
UNIT 2

• Low-Resource NLP

• Summarization

• Dialogue

• Question Answering

• Commonsense Reasoning
AI vs. Human Intelligence

Microsoft's Deep Learning Project Outperforms Humans In Image Recognition
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.

Microsoft claims new speech recognition record, achieving a super-human 5.1% error rate
New AI Model Exceeds Human Performance at Question Answering

2018

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

2020
Is AI a solved problem?
What do we know so far?

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

\[ x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) \]

- “panda” 57.7% confidence
- “nematode” 8.2% confidence
- “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
Kahneman’s “three cognitive systems”

Where are we and where do we go

Kahneman’s “three cognitive systems”


- Perception
- Intuition System 1
- Reasoning System 2

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”


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”


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

- Symbolic commonsense graph
- Neural commonsense representations
- Reasoning engine with common sense
- Constructing challenge datasets right
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

Language Model
Parrots are among the most intelligent birds, and the ability of some species to imitate human speech enhances their popularity as ___ → pets

Masked Language Model
Parrots are among the most intelligent [MASK], and the ability of some species to imitate human speech enhances their popularity as pets.
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):
  o BERT
  o Hand-crafted templates scored by GPT2
  o ConceptNet, mining from Wikipedia
  o Performs worse than supervised methods on ConceptNet but is more likely to generalize to different domains

<table>
<thead>
<tr>
<th>Candidate Sentence $S_i$</th>
<th>$\log p(S_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>“musician can playing musical instrument”</td>
<td>$-5.7$</td>
</tr>
<tr>
<td>“musician can be play musical instrument”</td>
<td>$-4.9$</td>
</tr>
<tr>
<td>“musician often play musical instrument”</td>
<td>$-5.5$</td>
</tr>
<tr>
<td>“a musician can play a musical instrument”</td>
<td>$-2.9$</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>Prompt</th>
<th>Model Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A ___ has fur.</td>
<td>dog, cat, fox, ...</td>
</tr>
<tr>
<td>A ___ has fur, is big, and has claws.</td>
<td>cat, bear, lion, ...</td>
</tr>
<tr>
<td>A ___ has fur, is big, has claws, has teeth, is an animal, eats, is brown, and lives in woods.</td>
<td>bear, wolf, cat, ...</td>
</tr>
</tbody>
</table>

- 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

\[ \text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{1}{\text{rank}_i} \]

\(\text{rank}_i\) is rank position of the first relevant document for the \(i\)-th query,
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)

<table>
<thead>
<tr>
<th>Context</th>
<th>Human Response</th>
<th>Human PF</th>
<th>ROBERTA-L Response</th>
<th>pLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Everyone knows that) a bear has ___ .</td>
<td>fur</td>
<td>27</td>
<td>teeth</td>
<td>.36</td>
</tr>
<tr>
<td></td>
<td>claws</td>
<td>15</td>
<td>claws</td>
<td>.18</td>
</tr>
<tr>
<td></td>
<td>teeth</td>
<td>11</td>
<td>eyes</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>cubs</td>
<td>7</td>
<td>ears</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>paws</td>
<td>7</td>
<td>horns</td>
<td>.02</td>
</tr>
<tr>
<td>(Everyone knows that) a ladder is made of ___ .</td>
<td>metal</td>
<td>25</td>
<td>wood</td>
<td>.33</td>
</tr>
<tr>
<td></td>
<td>wood</td>
<td>20</td>
<td>steel</td>
<td>.08</td>
</tr>
<tr>
<td></td>
<td>plastic</td>
<td>4</td>
<td>metal</td>
<td>.07</td>
</tr>
<tr>
<td></td>
<td>aluminum</td>
<td>2</td>
<td>aluminum</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>rope</td>
<td>2</td>
<td>concrete</td>
<td>.03</td>
</tr>
</tbody>
</table>

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

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

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

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
B. uses a hose to keep it from
C. gets the dog wet, then it
D. gets into a bath tub with

To separate egg whites from the yolk using a water bottle, you should...

- **Squeeze** the water bottle and press it against the yolk. **Release**, which creates suction and lifts the yolk.
- **Place** the water bottle and press it against the yolk. **Keep pushing**, which creates suction and lifts the yolk.
Symbolic reasoning 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

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

Talmor et al. 2020
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

<table>
<thead>
<tr>
<th></th>
<th>RoBERTa Large</th>
<th>BERT WWM</th>
<th>BERT Large</th>
<th>RoBERTa Base</th>
<th>BERT Base</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ALWAYS-NEVER</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AGE COMPARISON</strong></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>OBJECTS COMPAR.</strong></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ANTONYM NEG.</strong></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>PROPERTY CONJ.</strong></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TAXONOMY CONJ.</strong></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td><strong>ENCYC. COMP.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>MULTI-HOP COMP.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

Talmor et al. 2020
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

Abductive NLI
Physical IQa
Social IQa
VCR
HELLA SWAG

https://leaderboard.allenai.org/
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

<table>
<thead>
<tr>
<th></th>
<th>Small scale</th>
<th>Large scale</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Creation</strong></td>
<td>Expert-curated</td>
<td>Crowdsourced/automatic</td>
</tr>
<tr>
<td><strong>Coverage</strong></td>
<td>Limited coverage</td>
<td>Large coverage</td>
</tr>
<tr>
<td><strong>Training</strong></td>
<td>Dev/test only</td>
<td>Training/dev/test</td>
</tr>
<tr>
<td><strong>Budget</strong></td>
<td>Expert time costs</td>
<td>Crowdsourcing costs</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Small scale</th>
<th>Large scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creation</td>
<td>Expert-curated</td>
<td>Crowdsourced/automatic</td>
</tr>
<tr>
<td>Coverage</td>
<td>Limited coverage</td>
<td>Large coverage</td>
</tr>
<tr>
<td>Training</td>
<td>Dev/test only</td>
<td>Training/dev/test</td>
</tr>
<tr>
<td>Budget</td>
<td>Expert time costs</td>
<td>Crowdsourcing costs</td>
</tr>
</tbody>
</table>

**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?
Alex spilt food all over the floor and it made a huge mess. What will Alex want to do next?

**What happens next?**

- 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.
• Seemingly “super-human” performance by large pretrained LMs (BERT, GPT, etc.)

Benchmark creation important to avoid overstating performance (“super-human machine”)
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]

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

Personal experiences [Conway et al., 2000]

World knowledge and commonsense [Kintsch, 1988]
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.
Overview of existing resources

- **Open Mind Common Sense** (Minsky, Singh & Havasi, 1999)
- **Cyc** (Lenat et al., 1984)
- **ConceptNet** (Liu & Singh, 2004)
- **OpenCyc** (Lenat, 2004)
- **ResearchCyc** (Lenat, 2006)
- **OpenCyc 4.0** (Lenat, 2012)
- **Web Child** (Tandon et al., 2014)
- **Web Child 2.0** (Tandon et al., 2017)
- **NELL** (Carlson et al., 2010)
- **NELL** (Mitchell et al., 2015)
- **ConceptNet 5.5** (Speer et al., 2017)
- **ATOMIC** (Sap et al., 2019)

*today*
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

Representation

- Symbolic
- Natural language

Knowledge type

- Domain-specific
- Semantic
- Inferential
CONCEPTNET:

*semantic knowledge in natural language form*

http://conceptnet.io/
**reading**

An English term in ConceptNet 5.8

**Related terms**
- book
- books
- book

**Effects of reading**
- learning
- ideas
- a headache

**reading is a subevent of...**
- you learn
- turning a page
- learning

**Subevents of reading**
- relaxing
- study
- studying for a subject

**Things used for reading**
- article
- a library
- literature
- a paper page

**Types of reading**
- browse (n, communication)
- bumf (n, communication)
- clock time (n, time)
- 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

- X wanted to protect others
- X wanted to save themselves
- X wants to leave the scene
- X feels angry
- X feels tired
- X wants to file a police report
- X's heart races
- X needs to train hard
- X needs to know self-defense
- X is skilled
- X is brave
- X is strong
- X is seen as

- as a result, X wants
- has an effect on X
- has an effect on Y

- as a result, Y feels
- Y feels weak
- Y feels ashamed
- Y wants to run home
- Y wants to attack X again
- Y gets hurt
- Y falls back
X repels Y's attack

Before, X needed to train hard because X wanted to protect others. X needed to know self-defense before, X is seen as X is skilled, X is strong, X is brave.

As a result, X wants to save themselves, X wants to leave the scene, X wants to file a police report. As a result, X feels angry, X feels tired, X's heart races, X gains an enemy.

As a result, Y feels weak, Y feels ashamed, Y wants to run home, Y wants to attack X again. Y feels the effect on Y, X feels the effect on X. Y gets hurt, Y falls back.

Nine inference dimensions
X repels Y's attack
300,000 event nodes to date
880,000 if-Event-then-* knowledge triples
Overview of existing resources

- **Open Mind Common Sense** (Singh, 2002)
- **Cyc** (Lenat et al., 1984)
- **ConceptNet** (Liu & Singh, 2004)
- **OpenCyc** (Lenat, 2004)
- **ResearchCyc** (Lenat, 2006)
- **OpenCyc 4.0** (Lenat, 2012)
- **Web Child** (Tandon et al., 2014)
- **Web Child 2.0** (Tandon et al., 2017)
- **NELL** (Carlson et al., 2010)
- **NELL** (Mitchell et al., 2015)
- **ConceptNet 5.5** (Speer et al., 2017)
- **ATOMIC** (Sap et al., 2019)
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)

- **NELL**
  (Mitchell et al., 2015)
- **OpenCyc 4.0**
  (Lenat, 2012)

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

- **ConceptNet 5.5**
  (Speer et al., 2017)
- **ATOMIC**
  (Sap et al., 2019)

$(\text{\#implies} \ (\text{\#and} \ 
(\text{\#isa} \ ?OBJ \ ?SUBSET) \\
(\text{\#genls} \ ?SUBSET \ ?SUPERSET)) \\
(\text{\#isa} \ ?OBJ \ ?SUPERSET))$
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

ATOMIC (Sap et al., 2019)

- to be safe
- to protect himself
- to protect themselves
- to defend themselves
- to defend himself

**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.

WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale. Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. AAAI 2020
Katrina had the financial means to afford a new car while Monica did not, since [MASK] had a high paying job.
Static KB integration

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

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

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. [...]
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)
Knowledge Sources

Knowledge Bases
- WordNet
- SentiWordNet
- ATOMIC

Mining from Text
- Mining script knowledge from corpora, event plausibility from corpora

Tools
- Knowledge base embeddings, sentiment analysis models, COMET
[CLS] Katrina had the financial means to afford a new car while Monica did not, since [SEP] Katrina had a high paying job.
Combination Method

1. Incorporate into scoring function

2. Symbolic $\rightarrow$ vector representation
   ○ (+attention)

3. Multi-task learning
Combination Method

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

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

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

(Boschut et al., 2019)
Going beyond KBs

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

Generate dynamic graph with COMET

- X keeps X's temper
- X wants to show
- X avoids a fight
- X keeps X's in check
- no linking
- Kai stays calm
- Kai is viewed as cautious
- Kai wants to avoid trouble
- Kai intends
- context-free 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?