Commonsense Knowledge Acquisition and Reasoning

Yangqiu Song
Department of CSE, HKUST

Special thanks to Tianqing Fang, Zizheng Lin, Hongming Zhang for their contribution of slides.
Understanding Human’s Language Requires Complex Knowledge

• "Crucial to comprehension is the knowledge that the reader brings to the text. The construction of meaning depends on
  • the reader's knowledge of the language,
  • the structure of texts, a knowledge of the subject of the reading,
  • and a broad-based background or world knowledge.” (Day and Bamford, 1998)

• Contexts and knowledge contributes to the meanings

https://www.thoughtco.com/world-knowledge-language-studies-1692508
An Example of NLP

A dog is chasing a boy on the playground.

Lexical analysis (part-of-speech tagging)

Semantic analysis

Dog(d1).
Boy(b1).
Playground(p1).
Chasing(d1,b1,p1).

Syntactic analysis (Parsing)

A person saying this may be reminding another person to get the dog back...

Inference

Scared(b1) if Chasing(_,x,)._.

Pragmatic analysis (speech act)

Slides from Chengxiang Zhai and Hongning Wang
Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining By ChengXiang Zhai, Sean Massung
The State of the Art

A dog is chasing a boy on the playground

Semantics: some aspects
- Entity/relation extraction
- Word sense disambiguation
- Anaphora resolution

Inference: ???

Speech act analysis: ???

POS Tagging: 97%
Parsing: 90% on WSJ

Speech act analysis: ???

Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining By ChengXiang Zhai, Sean Massung
Pragmatics - Implicature

• “An implicature is something the speaker suggests or implies with an utterance, even though it is not literally expressed.” (Wikipedia)

A: What are they doing?
B: The firefighters should move the ____ quickly.

- boy/cat.
- rock.

• Relevant world knowledge
  - There is probably a fire engine around.
  - They are probably geared up.
  - There maybe other people looking at them.

• More ignorable commonsense
  - Firefighters are rescuers.
  - Firefighters are human beings.
  - There are more than one person.

• There is someone/something in danger.
• They are cooperating to save (the case).
“Commonsense Knowledge”

• When we communicate,
  • we omit a lot of “common sense” knowledge, which we assume the hearer/reader possesses
  • we keep a lot of ambiguities, which we assume the hearer/reader knows how to resolve

• A lemon is sour.
  • Attributes of objects
• To open a door, you must usually first turn the doorknob.
  • Condition/consequence of actions
• If you forget someone’s birthday, they may be unhappy with you.
  • Cause/effect between events and states

• Social:
  • If you forget your friend’s birthday, he/she may be mad at you.
• Physical:
  • Apples fall instead of floating in the air.
• World Entities:
  • Lions are bigger than cats.
In this tutorial, I will introduce

• How to collect commonsense knowledge? (Part 1)

• What we can do so far for commonsense reasoning and related tasks? (Part 2)
How to Collect Commonsense Knowledge?

• Motivation

• Information Extraction
How to Define Commonsense Knowledge as Computer Scientists? (Liu & Singh, 2004)

• “While to the average person the term ‘commonsense’ is regarded as synonymous with ‘good judgement’,”

• “the AI community it is used in a technical sense to refer to the millions of basic facts and understandings possessed by most people.”

• “Such knowledge is typically omitted from social communications”, e.g.,
  • If you forget someone’s birthday, they may be unhappy with you.

H Liu and P Singh, ConceptNet - a practical commonsense reasoning tool-kit, BTTJ, 2004
ConceptNet: An Approach Developed 16 Years Ago

• ConceptNet5 (Speer and Havasi, 2012)
  • Core is from Open Mind Common Sense (OMCS) (Liu & Singh, 2004)

Essentially a crowdsourcing based approach + text mining
ATOMIC: Everyday If-then Commonsense Knowledge

• These are day-to-day knowledge that help us understand each other.

• If a person $X$ did something, human beings are able to inference:
  • Motivation: Why person $X$ did this.
  • Pre-conditions: What enables $X$ to do this.
  • Characteristics: What are attributes of $X$.
  • Result: What will affect $X$/others

ATOMIC: Everyday If-then Commonsense Knowledge

• Define 4 categories of if-then relations:
  • Causes-agent (Motivation & Pre-condition): xIntend, xNeed
  • Stative (Characteristics): xAttr
  • Effects-agent (Results on X): xWant, xReact, xEffect
  • Effects-theme (Results on others): oWant, oReact, oEffect

Crowdsourcing 9 Types of IF-THEN relations

Arbitrary texts: Human annotation

All personal entity information has been removed to reduce ambiguity

Ways of Collecting Commonsense Knowledge

• Crowdsourcing
  • Pros
    • High quality
    • With proper quality control
    • Human can be creative when writing answers
    • Reflecting the ambiguity of language use
  • Cons
    • Ways of collection will limit the objects
      • Training Turk users: overfitting to the supervisor?
      • Time and money cost
    • Difficult to make the careful distinctions in quantifier structure
    • When used to train a machine learning algorithm
      • Selection bias

• Information extraction
  • Pros
    • Large-scale free text to use
    • Automatic and low time/money cost
    • Better coverage of more objects to reflect the world knowledge
  • Cons
    • Reporting bias
      • Frequency may not reflect preference
    • Rules may be inadequate
    • Noisy data
    • Lack of principles to perform extraction

How about a combination of two approaches?
• Accurate annotation (KB1)
• Automatic extraction + conceptualization and generation (KB2)
• Learning to populate KB1 with KB2 if they share similar structure

In fact, different commonsense knowledge bases have different properties
Revisit the Correlations of Selectional Preference and OMCS (ConceptNet)

(sing, song) (dobj, 9.25)
(song, UsedFor, sing)

(phone, ring) (nsubj, 8.75)
(phone, CapableOf, ring)

(cold, water) (amod, 8.86)
(water, HasProperty, cold)

(create, new) (dobj_amod, 8.25)
(create idea, UsedFor, invent new things)

(hungry, eat) (nsubj_amod, 10.00)
(eat, MotivatedByGoal, are hungry)
Transform ASER to ATOMIC
Coverage and Implicit Edges

- Most event related commonsense relations are implicit on ASER
- ConceptNet (Event-related relations), ATOMIC, ATOMIC 2020, and GLUCOSE

<table>
<thead>
<tr>
<th></th>
<th>$\text{ASER}_{\text{norm}}$ Coverage</th>
<th>Avg. Degree in $\text{ASER}_{\text{norm}}$</th>
<th>Avg. Degree in $\text{C}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>head(%)</td>
<td>tail(%)</td>
<td>edge(%)</td>
</tr>
<tr>
<td>ATOMIC</td>
<td>79.76</td>
<td>77.11</td>
<td>59.32</td>
</tr>
<tr>
<td>ATOMIC$_{20}$</td>
<td>80.39</td>
<td>47.33</td>
<td>36.73</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>77.72</td>
<td>54.79</td>
<td>43.51</td>
</tr>
<tr>
<td>GLUCOSE</td>
<td>91.48</td>
<td>91.85</td>
<td>81.01</td>
</tr>
</tbody>
</table>

Table 3: The overall matching statistics for the four CSKBs. The "edge" column indicates the proportion of edges where their heads and tails can be connected by paths in ASER. Average (in and out)-degree on $\text{ASER}_{\text{norm}}$ and $\text{C}$ for nodes from the CSKBs is also presented. The statistics in $\text{C}$ is different from (Malaviya et al., 2020) as we check the degree on the aligned CSKB $\text{C}$ instead of each individual CSKB.

So Far We Know That

• Some commonsense may appear in selectional preference when we talk

• Event and casual relations: explicit extraction may not be useful for commonsense
  • More inference and/or reasoning have to be performed

• How about language models?
Do Language Models Know Commonsense?

Sentence

If you forget someone’s birthday, they may be [MASK] with you.

Run Model

Model Output

Mask 1

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>If you forget someone’s birthday, they may be angry with you.</td>
<td>40.2%</td>
</tr>
<tr>
<td>If you forget someone’s birthday, they may be upset with you.</td>
<td>10.6%</td>
</tr>
<tr>
<td>If you forget someone’s birthday, they may be furious with you.</td>
<td>8.3%</td>
</tr>
<tr>
<td>If you forget someone’s birthday, they may be disappointed with you.</td>
<td>7.1%</td>
</tr>
<tr>
<td>If you forget someone’s birthday, they may be annoyed with you.</td>
<td>2.9%</td>
</tr>
</tbody>
</table>

https://demo.allennlp.org/masked-lm
GPT-2

Sentence
If you forget someone's birthday,

Run Model

Model Output

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>If you forget someone's birthday, you can tell them it ...</td>
<td>98.1%</td>
</tr>
<tr>
<td>If you forget someone's birthday, or you're confused or ...</td>
<td>1.4%</td>
</tr>
<tr>
<td>If you forget someone's birthday, let's change it for ...</td>
<td>0.5%</td>
</tr>
<tr>
<td>If you forget someone's birthday, the customer will be left ...</td>
<td>0%</td>
</tr>
<tr>
<td>If you forget someone's birthday, the cheque is not ...</td>
<td>0%</td>
</tr>
</tbody>
</table>

https://demo.allennlp.org/next-token-lm
To open a door, you must usually first turn the [MASK].

**BERT**

**Model Output**

**Mask 1**

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>To open a door, you must usually first turn the <strong>knob</strong>.</td>
<td>69.6%</td>
</tr>
<tr>
<td>To open a door, you must usually first turn the <strong>key</strong>.</td>
<td>11.9%</td>
</tr>
<tr>
<td>To open a door, you must usually first turn the <strong>lock</strong>.</td>
<td>9.9%</td>
</tr>
<tr>
<td>To open a door, you must usually first turn the <strong>handle</strong>.</td>
<td>7.3%</td>
</tr>
<tr>
<td>To open a door, you must usually first turn the <strong>locks</strong>.</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

https://demo.allennlp.org/masked-lm
GPT-2

Sentence
To open a door, you must usually first

Run Model

Model Output

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>To open a door, you must usually first go to the door.</td>
<td>97%</td>
</tr>
<tr>
<td>To open a door, you must usually first listen for the sounds of ...</td>
<td>2.7%</td>
</tr>
<tr>
<td>To open a door, you must usually first void the door with the ...</td>
<td>0.3%</td>
</tr>
<tr>
<td>To open a door, you must usually first square the room with your ...</td>
<td>0%</td>
</tr>
<tr>
<td>To open a door, you must usually first connect the pipes and doors ...</td>
<td>0%</td>
</tr>
</tbody>
</table>

https://demo.allennlp.org/next-token-lm
GPT-2

To open a door, you must usually first turn your head so that you ...

To open a door, you must usually first turn around and walk away.

To open a door, you must usually first turn to the left, through ...

To open a door, you must usually first turn around to get a grip ...

To open a door, you must usually first turn the small door open and ...

https://demo.allennlp.org/next-token-lm
### BERT

**Sentence**

A lemon is [MASK].

**Run Model**

---

**Model Output**

**Mask 1**

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A lemon is <strong>used</strong>.</td>
<td>18.9%</td>
</tr>
<tr>
<td>A lemon is <strong>eaten</strong>.</td>
<td>4.9%</td>
</tr>
<tr>
<td>A lemon is <strong>common</strong>.</td>
<td>4%</td>
</tr>
<tr>
<td>A lemon is <strong>preferred</strong>.</td>
<td>3.4%</td>
</tr>
<tr>
<td>A lemon is <strong>edible</strong>.</td>
<td>1.8%</td>
</tr>
</tbody>
</table>
**BERT**

Sentence

Lemon is [MASK].

**Model Output**

<table>
<thead>
<tr>
<th>Mask 1</th>
<th>Prediction</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemon is <strong>used</strong> .</td>
<td></td>
<td>7.7%</td>
</tr>
<tr>
<td>Lemon is <strong>eaten</strong> .</td>
<td></td>
<td>6.3%</td>
</tr>
<tr>
<td>Lemon is <strong>preferred</strong> .</td>
<td></td>
<td>6.3%</td>
</tr>
<tr>
<td>Lemon is <strong>common</strong> .</td>
<td></td>
<td>4.4%</td>
</tr>
<tr>
<td>Lemon is <strong>added</strong> .</td>
<td></td>
<td>2.4%</td>
</tr>
</tbody>
</table>

[https://demo.allennlp.org/masked-lm](https://demo.allennlp.org/masked-lm)
BERT

Sentence

A lemon is a [MASK].

Run Model

Model Output

Mask 1

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A lemon is a lemon.</td>
<td>9.9%</td>
</tr>
<tr>
<td>A lemon is a fruit.</td>
<td>5.8%</td>
</tr>
<tr>
<td>A lemon is a candy.</td>
<td>5%</td>
</tr>
<tr>
<td>A lemon is a dessert.</td>
<td>3.4%</td>
</tr>
<tr>
<td>A lemon is a plant.</td>
<td>2.4%</td>
</tr>
</tbody>
</table>

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

**Sentence**

Lemon is a [MASK].

**Run Model**

---

**Model Output**

### Mask 1

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemon is a <strong>nickname</strong>.</td>
<td>1.6%</td>
</tr>
<tr>
<td>Lemon is a <strong>synonym</strong>.</td>
<td>1.2%</td>
</tr>
<tr>
<td>Lemon is a <strong>surname</strong>.</td>
<td>1.2%</td>
</tr>
<tr>
<td>Lemon is a <strong>verb</strong>.</td>
<td>1%</td>
</tr>
<tr>
<td>Lemon is a <strong>pseudonym</strong>.</td>
<td>0.9%</td>
</tr>
</tbody>
</table>

[https://demo.allennlp.org/masked-lm](https://demo.allennlp.org/masked-lm)
BERT

Sentence

The taste of lemon is [MASK].

Run Model

Model Output

Mask 1

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>The taste of lemon is sweet.</td>
<td>30.4%</td>
</tr>
<tr>
<td>The taste of lemon is bitter.</td>
<td>17.4%</td>
</tr>
<tr>
<td>The taste of lemon is distinctive.</td>
<td>5.2%</td>
</tr>
<tr>
<td>The taste of lemon is unpleasant.</td>
<td>3.7%</td>
</tr>
<tr>
<td>The taste of lemon is pleasant.</td>
<td>2.9%</td>
</tr>
</tbody>
</table>

https://demo.allennlp.org/masked-lm
So Far We Know That

• Some commonsense may appear in selectional preference when we talk

• Event and casual relations: explicit extraction may not be useful for commonsense
  • More inference and/or reasoning have to be performed

• Large languages models probably need appropriate use (prompt) to get commonsense knowledge
How to Collect Commonsense Knowledge?

• Motivation

• Information Extraction
  • Do we have more principled ways of information extraction for commonsense knowledge?
• Knowledge in ConceptNet
  • Things
  • Spatial
  • Location
  • Events
  • Causal
  • Affective
  • Functional
  • Agents
Primitive Semantic Units in our Mind

- Semantic meaning in our language can be described as ‘a finite set of mental primitives and a finite set of principles of mental combination (Jackendoff, 1990)’.

- The primitive units of semantic meanings include
  - Thing (or Object),
  - Activity,
  - State,
  - Event,
  - Place,
  - Path,
  - Property,
  - Amount,
  - etc.

Knowledge Base

Traditional knowledge bases are mostly focused on entities/concepts and their attributes.

Slide Credit: Haixun Wang
Existing Knowledge Graphs

• Many large-scale knowledge graphs about entities and their attributes (property-of) and relations (thousands of different predicates) have been developed
  • Millions of entities and concepts
  • Billions of relationships

But how to characterize our mental world?
How to Grow a Mind?  
--Statistics, Structure, and Abstraction

• “In coming to understand the world—in learning concepts, acquiring language, and grasping causal relations—our minds make inferences that appear to go far beyond the data available.”

• The ability of performing powerful abstraction is the key

• The inference are usually probabilistic
“Concepts are the glue that holds our mental world together”
--Gregory L. Murphy, NYU

Typicality can be probabilistic: both are birds, but a “robin” is a more *typical* bird than a “penguin”
Why Are Concepts So Important?

• I steal several slides from Push Singh, the creator of OMCS and ConcepNet

---

Giving Computers Common Sense

Push Singh
MIT Media Lab
Common Sense Computing
9 February 2005

---

Our projects

• LifeNet (temporal probabilistic model)
• ConceptNet (large-scale semantic net)
• StoryNet (structured story knowledge base)
• GoalNet (typical human goals and priorities)
• SituationNet (prototypical situations)
• ShapeNet (shape kb for visual commonsense)
• GlueNet (connecting representations)
• ThinkNet (reflective reasoning with stories)
• ComicKit (telling stories by writing online comics)
• Serendipity (learning behavior from experience)
• ConceptMiner (terascale web mining)
• EM-ONE (implementing the Emotion Machine)

---

Representing Knowledge in Multiple Ways

Representing Knowledge in Multiple Ways

**StoryNet**

- I felt bored
- I decided to go the beach
- I got in my car
- I drove to the beach
- It was very crowded
- I swam in the surf
- I got tired

**LifeNet**

- Story-like scripts
- Transframes
- Frame-arrays
- Picture-frames

**ShapeNet**

- Semantic nets
- K-lines
- Neural nets

**ConceptNet**

- Micronemes

**SituationNet**

- Procedural, linguistic, physical, social, visual, haptic, among others

Representing Knowledge in Multiple Ways

Representing Knowledge in Multiple Ways

ShapeNet: Spatial Common Sense

Representing Knowledge in Multiple Ways

Representing Knowledge in Multiple Ways

SituationNet: detailed descriptions of situations

Prototypical situations

“When you get an idea and want to “remember” it, you create a K-line for it.”

“When later activated, the K-line induces a partial mental state resembling the partial mental state that created that K-line.”

“A partial mental state is a subset of those mental agencies operating at one moment.”

Representing Knowledge in Multiple Ways

• Encode memories in “abstract” form.
• Search all memory for the “nearest match.”
• Use prototypes with detachable defaults.
• Remember “methods,” not “answers.”
  • To get the mind into the (partial) state that solve the old problem, and then the mind might be able to handle the new problem in “the same way”.

**Commonsense Reasoning**

- **Conceptualization** and its *compositionality* in a sentence is one of the keys to commonsense reasoning (generalization), but there is still lack of study.

The CSKB is usually incomplete. So there is no direct support to entail the conclusion Y. Simple similarity/analogy does not always work, especially when training data is small (see Winograd Schema Challenge and Winogrande).

**CSKB/Training Data**
- Computer *not fit in* parcel, *REASON*, Computer *is big*
- Rock *not fit in* carrier, *REASON*, rock *is big*
- ...

**Induction**

- **X:** *Item does not fit in* container, *REASON*, item *is big*
- Trophy *is an item*; Suitcase *is a container*

**Deduction**

- **Y:** *Trophy does not fit in* suitcase, *REASON*, it *is big*
- **Consistent**
- **If we instantiate all, it’s possible to entail**

**Grounding**

- **Conceptualization**

Current deep learning models do not perform concept-level induction. Instead, they use model induction to summarize all they observe in the training data. That also means, they conceptualization ability is restricted to what they have seen.
Commonsense Reasoning

• The other way of doing conceptualization cannot help reasoning;
• Simple similarity does not explain this error.

PersonX eats cookies, xWant, to get some milk

to get some beverage

to get some dairy product
The K-Line Theory

• Attach a K-node (a mental state, KE) to a “Pyramid” agent (PE) at a certain level
  • The pyramid is a tree structure that we conceptualize the world
  • The mapping has a lower-band limit and a higher band limit, to compare the right common, non-conflicting properties
    • E.g., mapping Tesla to a company, big company, IT company, AI company, high-tech company, automobile company, when comparing it with Google, Toyota, some small company, needs the right level of comparison

• Then the partial states in PE will help us to make abstraction, logical and procedural reasoning
  • A lower K-line could affect the instantiation of a higher-level, “more abstract” K-line

X: Item does not fit in container, REASON, item is big
  Trophy is an item; Suitcase is a container

Conceptualization

Instantiation

Y: Trophy does not fit in suitcase, REASON, it is big

Representing Knowledge in Multiple Ways

• This is why we are building the concept-level representations of events

Before talking about ASER, we need to find a knowledge base for conceptualization

ProBase

A Probabilistic Knowledge Base

1. More than 2.7 million concepts automatically harnessed from 1.68 billion documents

2. Computation/Reasoning enabled by scoring:
   - **Consensus:** e.g., is there a company called Apple?
   - **Typicality:** e.g., how likely you think of Apple when you think about companies?
   - **Ambiguity:** e.g., does the word Apple, sans any context, represent Apple the company?
   - **Similarity:** e.g., how likely is an actor also a celebrity?
   - **Freshness:** e.g., *Pluto as a dwarf planet* is a claim more fresh than *Pluto as a planet.*

3. Give machines a new CPU (Commonsense Processing Unit) powered by a distributed graph engine called Trinity.

4. A little knowledge goes a long way after machines acquire a human touch.
Data Sources

• Patterns for single statements
  • Concept-instance “IsA” relationship: Hearst pattern [Hearst, 1992] (“A such as B, C and D”, etc.)
    • Good: “countries such as USA and Japan ...”
    • Tough: “animals other than cats such as dogs ...”
  • Handling multi-word expressions:
    • “domestic animals such as cats and dogs ...”
  • Instance-attributes: “What is A of B?”, etc.

• Semantic cleaning
  • Mutual exclusive

• Machine learning (e.g., Yu et al., 2020)
  • May Improve recall but reduce accuracy
  • Still working on single word concepts (mention detection is a big problem)
Probase is a large, universal, probabilistic knowledge base with an extremely large concept space.

Data are available at https://concept.research.microsoft.com/
Slide Credit: Haixun Wang
Nodes: Concepts

Probase:
- 2.7 M concepts automatically harnessed

Freebase:
- 2 K concepts built by community effort

Cyc:
- 120 K concepts
  - 25 years human labor

Slide Credit: Haixun Wang
Conceptualization with \textbf{ProBase}

Typicality \[ P(\text{concept} \mid \text{instance}) = \frac{\#(\text{concept, instance})}{\#(\text{instance})} \]

- **Robin**
  - bird
  - species
  - character
  - songbird
  - common bird
  - small bird

- **Penguin**
  - animal
  - bird
  - species
  - flightless bird
  - seabird
  - diving bird

Data are available at [https://concept.research.microsoft.com/](https://concept.research.microsoft.microsoft.com/)
Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, Weizhu Chen: Short Text Conceptualization Using a Probabilistic Knowledgebase. IJCAI 2011: 2330-2336
Primitive Semantic Units in our Mind

• Semantic meaning in our language can be described as ‘a finite set of mental primitives and a finite set of principles of mental combination (Jackendoff, 1990)’.

• The primitive units of semantic meanings include
  • Thing (or Object),
  • Activity,
  • State,
  • Event,
  • Place,
  • Path,
  • Property,
  • Amount,
  • etc.

Semantic Primitive Units

• Entities or concepts can be nouns or noun phrases
  • Concepts in ProBase (2012):
    • Company,
    • IT company,
    • big company,
    • big IT company,
    • ...  
  • Hierarchy is partially based on head+modifier composition
    • Noun + noun: e.g., IT company
    • Adj + noun: e.g., big company

• Let’s think about verbs and verb phrases
  • How should we define semantic primitive unit for verbs?
“Linguistic Description – Grammar = Semantics”
The lower bound of a semantic theory (Katz and Fodor, 1963)

- Disambiguation needs both “the speaker's knowledge of his language and his knowledge about the world” (Katz and Fodor, 1963)
  - The bill is large.
    - Some document demanding a sum of money to discharge a debt exceeds in size most such documents
    - The beak of a certain bird exceeds in bulk those of most similar birds
  - Syntactically unambiguous
  - Compare semantic meanings by fixing grammar

Selectional Preference (SP)

- The need of language inference based on ‘partial information’ (Wilks, 1975)
  - The soldiers fired at the women, and we saw several of them fall.
  - The needed partial information: hurt things tending to fall down
    - “not invariably true”
    - “tend to be of a very high degree of generality indeed”

- Selectional preference (Resnik, 1993)
  - A relaxation of selectional restrictions (Katz and Fodor, 1963) and as syntactic features (Chomsky, 1965)
  - Applied to isA hierarchy in WordNet and verb-object relations

A Test of Commonsense Reasoning

• Proposed by Hector Levesque at U of Toronto
• An example taking from Winograd Schema Challenge

  • (A) The fish ate the worm. It was hungry.
  • (B) The fish ate the worm. It was tasty.

• On the surface, they simply require the resolution of anaphora
  • But Levesque argues that for Winograd Schemas, the task requires the use of knowledge and commonsense reasoning
Why is it a challenge?

• Must also be carefully written not to betray their answers by selectional restrictions or statistical information about the words in the sentence

• Designed to be an improvement on the Turing test
A Brief History of Datasets and Development

The first large dataset.
Rahman and Ng: EMNLP-CoNLL 2012
Stanford: 55.19%
Their system: 73.05%

"Strictly speaking, we are addressing a relaxed version of the Challenge: while Levesque focuses solely on definite pronouns whose resolution requires background knowledge not expressed in the words of a sentence, we do not impose such a condition on a sentence."

Levesque. AAAI Spring Symposium 2011
Davis et al. "A Collection of Winograd Schemas" 2014
The first round of the challenge was a collection of 60 Pronoun Disambiguation Problems (PDPs). The highest score achieved was 58% correct, by Quan Liu, from University of Science and Technology, China.

<table>
<thead>
<tr>
<th>Author/year</th>
<th>System</th>
<th>Fine-tuned</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emami et al. (2018)</td>
<td>Knowledge Hunter</td>
<td>No</td>
<td>54.58%</td>
</tr>
<tr>
<td>Trieu H. Trinh and Quoc V. Le (2018)</td>
<td>Language models (single)</td>
<td>No</td>
<td>54.58%</td>
</tr>
<tr>
<td></td>
<td>Language models (Ensemble)</td>
<td>No</td>
<td>63.74%</td>
</tr>
<tr>
<td>Alec Radford et al. (2019)</td>
<td>GPT-2</td>
<td>No details</td>
<td>70.70%</td>
</tr>
<tr>
<td>Ruan et al. (2019)</td>
<td>BERT-large + dependency</td>
<td>Rahman and Ng 2012 dataset</td>
<td>71.10%</td>
</tr>
<tr>
<td>Kocijan et al. (2019)</td>
<td>BERT-large</td>
<td>No</td>
<td>60.10%</td>
</tr>
<tr>
<td></td>
<td>GPT</td>
<td>No</td>
<td>55.30%</td>
</tr>
<tr>
<td></td>
<td>Wiki + Rahman and Ng 2012 dataset</td>
<td></td>
<td>72.20%</td>
</tr>
</tbody>
</table>

- Human’s performance: 92.1% (Bender 2015)
- WinoGrande (RoBERTa + 43K Training data): 90.1% (Sakaguchi et al., 2019)
SP-10K: A Large-scale Evaluation Set

- Traditional evaluation
  - Small sets of one-hop direct dependency relations
    - McRae et al., 1998: 821 pairs of nsubj and dobj relations
    - Keller and Lapata, 2003: 540 pairs of dobj, noun-noun, and amod relations
    - Padó et al., 2006: 207 pairs of nsubj, dobj, and amod relations
    - Wang et al, 2018: 3062 (subject, verb, dobject) triplets
  - Pseudo-disambiguation (Ritter et al., 2010; de Cruys, 2014): corpus driven, no human annotation

- Ours:
  - 10K pairs of five relations, including two 2-hop relations
## Examples in SP-10K

<table>
<thead>
<tr>
<th>dobj</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(eat, meal)</td>
<td>10.00</td>
</tr>
<tr>
<td>(close, door)</td>
<td>8.50</td>
</tr>
<tr>
<td>(touch, food)</td>
<td>5.50</td>
</tr>
<tr>
<td>(hate, invest)</td>
<td>4.00</td>
</tr>
<tr>
<td>(eat, mail)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nsubj</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(singer, sing)</td>
<td>10.00</td>
</tr>
<tr>
<td>(law, permit)</td>
<td>7.78</td>
</tr>
<tr>
<td>(women, pray)</td>
<td>5.83</td>
</tr>
<tr>
<td>(victim, contain)</td>
<td>2.22</td>
</tr>
<tr>
<td>(textbook, eat)</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>amod</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(fresh, air)</td>
<td>9.77</td>
</tr>
<tr>
<td>(new, method)</td>
<td>8.89</td>
</tr>
<tr>
<td>(medium, number)</td>
<td>4.09</td>
</tr>
<tr>
<td>(immediate, food)</td>
<td>2.05</td>
</tr>
<tr>
<td>(secret, wind)</td>
<td>0.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dobj_amod</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(lift, heavy object)</td>
<td>9.17</td>
</tr>
<tr>
<td>(design, new object)</td>
<td>8.00</td>
</tr>
<tr>
<td>(attack, small object)</td>
<td>5.23</td>
</tr>
<tr>
<td>(inform, weird object)</td>
<td>3.64</td>
</tr>
<tr>
<td>(earn, rubber object)</td>
<td>0.63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nsubj_amod</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(evil subject, attack)</td>
<td>9.00</td>
</tr>
<tr>
<td>(recent subject, demonstrate)</td>
<td>6.00</td>
</tr>
<tr>
<td>(random subject, bear)</td>
<td>4.00</td>
</tr>
<tr>
<td>(happy subject, steal)</td>
<td>2.25</td>
</tr>
<tr>
<td>(sunny subject, make)</td>
<td>0.56</td>
</tr>
</tbody>
</table>
Correlations with OMCS

- (sing, song) (dobj, 9.25)
  (song, UsedFor, sing)
- (phone, ring) (nsubj, 8.75)
  (phone, CapableOf, ring)
- (cold, water) (amod, 8.86)
  (water, HasProperty, cold)
- (create, new) (dobj_amod, 8.25)
  (create idea, UsedFor, invent new things)
- (hungry, eat) (nsubj_amod, 10.00)
  (eat, MotivatedByGoal, are hungry)
Performance on Winograd Schema

• 72 out of 273 questions satisfying nsubj_amod and dobj_amod relations
  • Jim yelled at Kevin because he was so upset.
  • We compare the scores
    • (yell, upset object) following nsubj_amod
    • (upset object, yell) following dobj_amod

• Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Correct</th>
<th>Wrong</th>
<th>NA</th>
<th>Accuracy (predicted)</th>
<th>Accuracy (overall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford</td>
<td>33</td>
<td>35</td>
<td>4</td>
<td>48.5%</td>
<td>48.6%</td>
</tr>
<tr>
<td>End2end (Lee et al., 2018)</td>
<td>36</td>
<td>36</td>
<td>0</td>
<td>50.0%</td>
<td>50.0%</td>
</tr>
<tr>
<td>PP* (Resnik, 1997)</td>
<td>36</td>
<td>19</td>
<td>17</td>
<td>65.5%</td>
<td>61.8%</td>
</tr>
<tr>
<td>SP-10K</td>
<td>13</td>
<td>0</td>
<td>56</td>
<td>100%</td>
<td>59.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dobj_amod</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(lift, heavy object)</td>
<td>9.17</td>
</tr>
<tr>
<td>(design, new object)</td>
<td>8.00</td>
</tr>
<tr>
<td>(attack, small object)</td>
<td>5.23</td>
</tr>
<tr>
<td>(inform, weird object)</td>
<td>3.64</td>
</tr>
<tr>
<td>(earn, rubber object)</td>
<td>0.63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nsubj_amod</th>
<th>Plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>(evil subject, attack)</td>
<td>9.00</td>
</tr>
<tr>
<td>(recent subject, demonstrate)</td>
<td>6.00</td>
</tr>
<tr>
<td>(random subject, bear)</td>
<td>4.00</td>
</tr>
<tr>
<td>(happy subject, steal)</td>
<td>2.25</td>
</tr>
<tr>
<td>(sunny subject, make)</td>
<td>0.56</td>
</tr>
</tbody>
</table>

*PP: posterior probability for SP acquisition using Wikipedia data
KnowlyWood

• Perform information extraction from free text
  • Mostly movie scripts and novel books

• Four relations: previous, next, parent, similarity

• No subject information
  • Only verb+object

ASER (Activities, States, Events, and their Relations)

Mourelatos’ taxonomy (1978)

- **State**: The air smells of jasmine.
- **Process**: It’s snowing.
- **Development**: The sun went down.
- **Punctual occurrence**: The cable snapped. He blinked. The pebble hit the water.

Bach’s taxonomy (1986)

- **Static states**: be in New York, love (one's cat);
- **Dynamic states**: sit, stand, drunk, present, sick;
- **Processes**: walk, push a cart, sleep;
- **Protracted events**: build (a cabin), eat a sandwich, polish a shoe, walk to Boston;
- **Culminations**: take off; arrive, leave, depart;
- **Happenings**: blink, flash, knock, kick, hit, pat, wink;

Eventualities

- Using patterns to collect partial information

- Six relations are also kept but treated as auxiliary edges
  - advmod,
  - amod,
  - nummod,
  - aux,
  - compound,
  - neg

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Code</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>n1-nsubj-v1</td>
<td>s-v</td>
<td>'The dog barks'</td>
</tr>
<tr>
<td>n1-nsubj-v1-dobj-n2</td>
<td>s-v-o</td>
<td>'I love you'</td>
</tr>
<tr>
<td>n1-nsubj-v1-xcomp-a</td>
<td>s-v-a</td>
<td>'He felt ill'</td>
</tr>
<tr>
<td>n1-nsubj-(v1-iobj-n2)-dobj-n3</td>
<td>s-v-o-o</td>
<td>'You give me the book'</td>
</tr>
<tr>
<td>n1-nsubj-a1-cop-be</td>
<td>s-be-a</td>
<td>'The dog is cute'</td>
</tr>
<tr>
<td>n1-nsubj-v1-xcomp-a1-cop-be</td>
<td>s-v-be-a</td>
<td>'I want to be slim'</td>
</tr>
<tr>
<td>n1-nsubj-v1-xcomp-n2-cop-be</td>
<td>s-v-be-o</td>
<td>'I want to be a hero'</td>
</tr>
<tr>
<td>n1-nsubj-v1-xcomp-v2-dobj-n2</td>
<td>s-v-v-o</td>
<td>'I want to eat the apple'</td>
</tr>
<tr>
<td>n1-nsubj-v1-xcomp-v2</td>
<td>s-v-v</td>
<td>'I want to go'</td>
</tr>
<tr>
<td>(n1-nsubj-a1-cop-be)-nmod-n2-case-p1</td>
<td>s-be-a-p-o</td>
<td>'It' cheap for the quality'</td>
</tr>
<tr>
<td>n1-nsubj-v1-nmod-n2-case-p1</td>
<td>s-v-p-o</td>
<td>'He walks into the room'</td>
</tr>
<tr>
<td>(n1-nsubj-v1-dobj-n2)-nmod-n3-case-p1</td>
<td>s-v-o-p-o</td>
<td>'He plays football with me'</td>
</tr>
<tr>
<td>n1-nsubjpass-v1</td>
<td>spass-v</td>
<td>'The bill is paid'</td>
</tr>
<tr>
<td>n1-nsubjpass-v1-nmod-n2-case-p1</td>
<td>spass-v-p-o</td>
<td>'The bill is paid by me'</td>
</tr>
</tbody>
</table>
Eventuality Relations

- 14 relations taking from CoNLL shared task
  - More frequent relations
- Less ambiguous connectives
  - ‘so that’ 31 times only in ‘Result’ relations
- Some are ambiguous
  - ‘while’: Conjunction 39 times, Contrast 111 times, Expectation 79 times, and Concession 85 times
- Classifiers trained on Penn Discourse Treebank (PDTB) (Prasad et al., 2007)

<table>
<thead>
<tr>
<th>Relation Type</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precedence</td>
<td>E1 before E2; E1, then E2; E1 till E2; E1 until E2</td>
</tr>
<tr>
<td>Succession</td>
<td>E1 after E2; E1 once E2</td>
</tr>
<tr>
<td>Synchronous</td>
<td>E1, meanwhile E2; E1 meantime E2; E1, at the same time E2</td>
</tr>
<tr>
<td>Reason</td>
<td>E1, because E2</td>
</tr>
<tr>
<td>Result</td>
<td>E1, so E2; E1, thus E2; E1, therefore E2; E1, so that E2</td>
</tr>
<tr>
<td>Condition</td>
<td>E1, if E2; E1, as long as E2</td>
</tr>
<tr>
<td>Contrast</td>
<td>E1, but E2; E1, however E2; E1, by contrast E2; E1, in contrast E2; E1, on the other hand, E2; E1, on the contrary, E2</td>
</tr>
<tr>
<td>Concession</td>
<td>E1, although E2</td>
</tr>
<tr>
<td>Conjunction</td>
<td>E1 and E2; E1, also E2</td>
</tr>
<tr>
<td>Instantiation</td>
<td>E1, for example E2; E1, for instance E2</td>
</tr>
<tr>
<td>Restatement</td>
<td>E1, in other words E2</td>
</tr>
<tr>
<td>Alternative</td>
<td>E1 or E2; E1, unless E2; E1, as an alternative E2; E1, otherwise E2</td>
</tr>
<tr>
<td>ChosenAlternative</td>
<td>E1, E2 instead</td>
</tr>
<tr>
<td>Exception</td>
<td>E1, except E2</td>
</tr>
</tbody>
</table>

A Running Example

Eventuality Extraction

Relation Extraction

Graph Construction

Conceptualization

An input sentence

My army will find your boat. In the meantime, I'm sure we could find you suitable accommodations.
Scales of Verb Related Knowledge Graphs

- FrameNet (Baker et al., 1998)
- ACE (Aguilar et al., 2014)
- PropBank (Palmer et al., 2005)
- TimeBank (Pustejovsky et al., 2003)
- OMCS in ConceptNet (Liu & Singh, 2004)
- Event2Mind (Smith et al., 2018)
- Propora (Dalvi et al., 2018)
- ATOMIC (Sap et al., 2018)
- Knowlywood (Tandon et al., 2015)
- ASER (core)
- ASER (full)

- #Eventualities
- #Relations

- 6000x larger
- 300x larger
So far we have:

• A concept based knowledge base: ProBase
  • There are many others
  • Hypernym detection is also an active research in NLP

• A verb-phrase based knowledge base: ASER

• How to conceptualize?
Inference for Winograd Schema Challenge

**Question**

97. *The fish* ate *the worm*. *It was hungry.*

98. *The fish* ate *the worm*. *It was tasty.*

**Extracted Eventualities**

- The fish: (‘X ate Y’, ‘X was hungry’)
- The worm: (‘X ate Y’, ‘Y was hungry’)
- The fish: (‘X ate Y’, ‘X was tasty’)
- The worm: (‘X ate Y’, ‘Y was tasty’)

**ASER Knowledge**

- ASER(‘X ate Y’, ‘X was hungry’) = 18
- ASER(‘X ate Y’, ‘Y was hungry’) = 1
- ASER(‘X ate Y’, ‘X was tasty’) = 0
- ASER(‘X ate Y’, ‘Y was tasty’) = 7

**Prediction**

- The fish
- the worm
Partial Information Aggregation

• “hurt things tending to fall down”

(hurt, X) connection (X, fall)

• “stocks price may increase when a company acquires a start-up”

(company, acquire, start-up) result-in (stock, increase)
Normalization

He, she, I, Bob, ... → PERSON 1.0
1996, 2020, 1949, ... → YEAR 1.0
23, 20, 333, ... → DIGIT 1.0
www.google.com, ... → URL 1.0
Conceptualization

\[
P(\text{ResultIn} | (\text{person, have, animal}), (\text{positive-emotion, come})) = 0.281 \times 3 \times 0.087 + 0.333 \times 2 \times 0.125 = 0.157
\]
Conceptualization Examples

Conceptualized ASER

- PersonX gives PersonY Red-Meat
  - Conjunction (0.05)

- PersonX be thirsty
  - Conjunction (1.0)
  - Precedence (0.057)
  - Succession (0.042)

- PersonX be hungry
  - Conjunction (0.077)
  - Precedence (0.042)
  - Succession (0.5)

- PersonX order Meat
  - Result (0.077)

- PersonX be full

Eventualities:
- He orders meat (Pr = 0.1)
- He orders beef (Pr = 0.2)
- He orders chicken (Pr = 0.1)

Graphs showing the relationship between extracted and conceptualized eventualities.

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, Yangqiu Song: ASER: Towards Large-scale Commonsense Knowledge Acquisition via Higher-order Selectional Preference over Eventualities. CoRR abs/2104.02137 (2021)
ASER 2.0

• 1.0 (in 2019): Rule based extraction (14 Eventuality Patterns, Improved Version)

<table>
<thead>
<tr>
<th>Data</th>
<th>#Unique Eventualities</th>
<th>#Unique Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>34 millions</td>
<td>15 millions</td>
</tr>
<tr>
<td>Full</td>
<td>272 millions</td>
<td>206 millions</td>
</tr>
</tbody>
</table>

• 2.0 (in 2021): Discourse Parser (18 Eventuality Patterns + Wang and Lan 2015)

<table>
<thead>
<tr>
<th>Data</th>
<th>#Unique Eventualities</th>
<th>#Unique Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>53 millions</td>
<td>52 millions</td>
</tr>
<tr>
<td>Full</td>
<td>439 millions</td>
<td>649 millions</td>
</tr>
</tbody>
</table>

• Conceptualization Core (Using top 5 concepts for each detected instance):
  • Concepts: 15 millions (based on 14 millions eventualities, 1.X times)
  • Concept Relations: 224 millions (based on 53 millions eventuality relations, 4.X times)
Rule Mining: Eventualities

• Mine Rules using AMIE +

\[ < E_a, T_1, E_b > \land < E_b, T_2, E_c > \Rightarrow < E_a, T_3, E_b >, \]

<table>
<thead>
<tr>
<th>Rule</th>
<th>Instances</th>
</tr>
</thead>
</table>
| \( \langle E_b \xrightarrow{\text{Concession}} E_f \rangle \land \langle E_a \xrightarrow{\text{Result}} E_f \rangle \Rightarrow \langle E_a \xrightarrow{\text{Contrast}} E_b \rangle \) | \( \langle \text{I do not know} \rightarrow \text{I guess} \rangle \land \langle \text{I believe} \rightarrow \text{I guess} \rangle \Rightarrow \langle \text{I believe} \rightarrow \text{I do not know} \rangle \)  
\( \langle \text{I am not sure} \rightarrow \text{I guess} \rangle \land \langle \text{I hope so} \rightarrow \text{I guess} \rangle \Rightarrow \langle \text{I hope so} \rightarrow \text{I am not sure} \rangle \)  
\( \langle \text{I understand} \rightarrow \text{I can not speak} \rangle \land \langle \text{I am not a lawyer} \rightarrow \text{I can not speak} \rangle \Rightarrow \langle \text{I am not a lawyer} \rightarrow \text{I understand} \rangle \) |
| \( \langle E_f \xrightarrow{\text{Contrast}} E_b \rangle \land \langle E_a \xrightarrow{\text{Instantiation}} E_f \rangle \Rightarrow \langle E_a \xrightarrow{\text{Contrast}} E_b \rangle \) | \( \langle \text{I remember} \rightarrow \text{I could not find it} \rangle \land \langle \text{I get} \rightarrow \text{I remember} \rangle \Rightarrow \langle \text{I get} \rightarrow \text{I could not find it} \rangle \)  
\( \langle \text{I would say} \rightarrow \text{I might be wrong} \rangle \land \langle \text{I hope} \rightarrow \text{I would say} \rangle \Rightarrow \langle \text{I hope} \rightarrow \text{I might be wrong} \rangle \)  
\( \langle \text{It have been suggested} \rightarrow \text{This is unlikely} \rangle \land \langle \text{It is possible} \rightarrow \text{It have been suggested} \rangle \Rightarrow \langle \text{It is possible} \rightarrow \text{This is unlikely} \rangle \) |
| \( \langle E_c \xrightarrow{\text{ChosenAlternative}} E_b \rangle \land \langle E_a \xrightarrow{\text{ChosenAlternative}} E_c \rangle \Rightarrow \langle E_a \xrightarrow{\text{ChosenAlternative}} E_b \rangle \) | \( \langle \text{I will not go} \rightarrow \text{You come here} \rangle \land \langle \text{I want to see} \rightarrow \text{I will not go} \rangle \Rightarrow \langle \text{I want to see} \rightarrow \text{You come here} \rangle \)  
\( \langle \text{I want} \rightarrow \text{It is} \rangle \land \langle \text{I wish} \rightarrow \text{I want} \rangle \Rightarrow \langle \text{I wish} \rightarrow \text{It is} \rangle \)  
\( \langle \text{I want} \rightarrow \text{I get} \rangle \land \langle \text{I do not get that} \rightarrow \text{I want} \rangle \Rightarrow \langle \text{I do not get that} \rightarrow \text{I get} \rangle \) |

Concession | E1, although E2

ChosenAlternative | E1, E2 instead

Rule Mining: Concepts

• Mine Rules using AMIE+

\[ <E_a, T_1, E_b> \land <E_b, T_2, E_c> \Rightarrow <E_a, T_3, E_b>, \]

<table>
<thead>
<tr>
<th>Rule</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restatement</td>
<td>(PersonX laugh (\rightarrow) PersonX smile) \land (PersonX laugh (\rightarrow) PersonX open Facial-Feature) (\Rightarrow) (PersonX smile (\rightarrow) PersonX open Facial-Feature)</td>
</tr>
<tr>
<td>Conjunction</td>
<td>(PersonX love it (\rightarrow) It be good) \land (PersonX love it (\rightarrow) It be tasty) (\Rightarrow) (It be good (\rightarrow) It be tasty)</td>
</tr>
<tr>
<td>Instantiation</td>
<td>(PersonX wish (\rightarrow) PersonX need) \land (PersonX wish (\rightarrow) PersonX need) (\Rightarrow) (PersonX need (\rightarrow) PersonX need)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rule</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concession</td>
<td>(PersonX know (\rightarrow) PersonX be sure) \land (PersonX know (\rightarrow) PersonX remember) (\Rightarrow) (PersonX be sure (\rightarrow) PersonX remember)</td>
</tr>
<tr>
<td>Contrast</td>
<td>(PersonX order Dish (\rightarrow) PersonX be hungry) \land (PersonX order Dish (\rightarrow) PersonX order) (\Rightarrow) (PersonX order (\rightarrow) PersonX be hungry)</td>
</tr>
</tbody>
</table>

**Instantiation**
- E1, **for example** E2; E1, **for instance** E2

**Restatement**
- E1, **in other words** E2

Incorporating More Relations

Two Issues:
1. Concept Transitivity
2. Verb’s Entailment Relations

Concept Graph

Eventuality Graph
Entailment Graph Construction

### Node Type

<table>
<thead>
<tr>
<th>Node Type</th>
<th>Reference</th>
<th>#Graphs</th>
<th>#Nodes</th>
<th>#Edges</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typed Predicate</td>
<td>Berant et al., ACL, 2011</td>
<td>2,303</td>
<td>10,672</td>
<td>263,756</td>
<td>Place/disease</td>
</tr>
<tr>
<td></td>
<td>Hosseini et al. TACL, 2018</td>
<td>363</td>
<td>101K</td>
<td>66M</td>
<td>News</td>
</tr>
<tr>
<td>Open IE Proposition</td>
<td>Levy et al., CoNLL, 2014</td>
<td>30</td>
<td>5,714</td>
<td>1.5M</td>
<td>Healthcare</td>
</tr>
<tr>
<td>Eventuality</td>
<td>Ours</td>
<td>473</td>
<td>10M</td>
<td>103M</td>
<td>Commonsense</td>
</tr>
</tbody>
</table>
Three-step Construction

Eventuality pre-processing

Local Inference

Global Inference
Other Resources

- ELG: An Event Logic Graph to discovery of evolutionary patterns among events
  - Sequential (the same meaning with “temporal”)
  - Causal
  - Conditional
  - Hypernym-hyponym (“is-a”) relations between events

- Causal Bank and Cause Effect Graph
  - Sentences expressing causal patterns
  - Lexical causal knowledge graph

Conclusions for This Part

• Commonsense has been a long standing core AI problem

• We have seen a sudden interest in commonsense recently

• We have talked about commonsense knowledge acquisition
  • Crowdsourcing
    • Learning upon annotated data will be introduced in the second part
  • Information extraction
    • How to formulate the problem
    • What have been done

• What’s missing?
  • We have done entity and eventuality based extraction
  • Other commonsense knowledge, e.g., physical knowledge, attribute (color, shape) knowledge were not mentioned
10 Minutes Break
In this tutorial, I will introduce

• How to collect commonsense knowledge? (Part 1)

• What we can do so far for commonsense reasoning and related tasks? (Part 2)
Learning and Reasoning with CSKB/CSKG

• Introduction

• Learning and Reasoning on CSKBs/CSKGs

• Learning and Reasoning for Downstream Tasks (CSQA)

Slides credit of this part: Tianqing Fang
Reasoning

• General reasoning
  • Logical reasoning: Given premise/presumption, draw conclusions based *solely* on the premise

• For example
  • $KB = \{Rain \rightarrow Wet, Rain\}, f = Wet$
  • Applying Modus ponens inference rule in KB:
    $$\frac{Rain, Rain \rightarrow Wet}{Wet}$$

Tell $f$  →  KB  →  ?

Already knew that: entailment $KB \models f$
Don't believe that: contradiction $KB \models \neg f$
Learned something new (update KB): contingent

Ask $f$  →  KB  →  ?

Yes: entailment $KB \models f$
No: contradiction $KB \models \neg f$
I don't know: contingent

Entailment $KB \models f$: KB defines more specific knowledge (configuration) than formula $f$, aka, $f$ added no information to KB
Commonsense Reasoning

- Commonsense reasoning in natural language:
  - Logical reasoning: E.g., first-order \texttt{IsA} relations. Taxonomy reasoning. (Davis 2017)
  - General natural language: Draw conclusions similar to humans’ \textit{folk psychology} and \textit{naive physics} (Davis 2015)

- Commonsense reasoning in traditional logics
  - Lacks such high-quality KB to perform logical reasoning
  - Can only deal with first-order logics like \texttt{IsA}
  - KB may be noisy. Needs probabilistic reasoning
  - Implicit inferential knowledge outside of the taxonomy

\[ \text{If } X \text{ hit } Y \text{ on the face, } Y \text{ will be (a) upset (b) happy} \]

Corgi is a kind of dog. Dog barks. --> Corgi barks.

Commonsense Reasoning

- **Conceptualization** and its **compositionality** in a sentence is one of the keys to commonsense reasoning (generalization), but there is still lack of study.

**CSKB/Training Data**
- Computer not fit in parcel, **REASON**, Computer is big
- Rock not fit in carrier, **REASON**, rock is big
- ...

The CSKB is usually incomplete. So there is no direct support to entail the conclusion Y. Simple similarity/analogy does not always work, especially when training data is small (see Winograd Schema Challenge and Winogradne).

- **Induction**
  - X: Item does not fit in container, **REASON**, item is big
  - Trophy is an item; Suitcase is a container

- **Grounding**
  - Computer not fit in parcel, **REASON**, Computer is big

- **Consistent**
  - Y: Trophy does not fit in suitcase, **REASON**, it is big

- **Instantiation**
  - If we instantiate all, it's possible to entail

- **Conceptualization**
  - Current deep learning models do not perform concept-level induction. Instead, they use model induction to summarize all they observe in the training data. That also means, they conceptualization ability is restricted to what they have seen.
Inference with Entailment

- Commonsense reasoning in current NLP community
  - Usually just textual entailment (learning an entailment classifier) and textual implication (Gordon et al. 2012)
    - “Entailment is meant to include inferences that are necessarily true due to the meaning of the text fragment.”
    - “Implications are inferences expected to be true, are likely causes or effects of the text, or are default assumptions”
  - Based not only on the context, but *world knowledge*
    - Able to leverage implicit knowledge using language models

Entailment can be done implicitly; this is why joint learning with NLP helps commonsense tasks

Reasoning Approaches and Typical Objectives (2015)

- **Reasoning architecture**: A closely related issue is the representation of the meaning of natural language sentences.
- **Plausible inference**: Drawing provisional or uncertain conclusions.
- **Range of reasoning modes**: Incorporating a variety of different modes of inference, such as explanation, generalization, abstraction, analogy, and simulation.
- **Painstaking analysis of fundamental domains**: Complex reasoning about basic domains such as time, space, naïve physics, and naïve psychology.
- **Breadth**: Attaining powerful commonsense reasoning will require a large body of knowledge.
- **Independence of experts**: Paying experts to hand-code a large knowledge base is slow and expensive.
- **Applications**: To be useful, the commonsense reasoner must serve the needs of applications and must interface with them smoothly.
- **Cognitive modeling**: Theories of commonsense automated reasoning accurately describe commonsense reasoning in people.

Learning and Reasoning with CSKB/CSKG

- Introduction

- Learning and Reasoning on CSKBs/CSKGs
  - Commonsense Knowledge Bases
  - Commonsense Knowledge Generation
  - Commonsense Knowledge Base Completion
  - Commonsense Knowledge Base Population

- Learning and Reasoning for Downstream Tasks (CSQA)

Slides credit of this part: Tianqing Fang
Commonsense Resources and Benchmarks

• The foundation of computational commonsense

• Why are Commonsense Knowledge Base (CSKB) needed
  • 60M knowledge about the world are needed (Marvin Minsky)
  • Commonsense is generally omitted in daily conversation
  • Commonsense knowledge is implicit knowledge that is hard to mine directly from existing corpora
  • Crowdsourcing is needed

Commonsense Knowledge Bases

• ConceptNet (v5.7)
  • Formalizing relations in OMCS and merge DBPedia, WordNet, etc.
  • Also incorporate multi-lingual word knowledge

Commonsense Knowledge Bases

• **ATOMIC**
  - Everyday If-then commonsense knowledge
  - Motivation, characteristics, and effects on agent/theme.

• **GLUCOSE**
  - Factors/emotions that enables/causes a event from stories.
    - grounded in narratives

If X hit Y on the face, Y will be upset

SomeoneA possesses Something Enables
SomeoneA moves it

## Commonsense Resources and Benchmarks

- **Scale and Comparisons of Large-scale CSKBs**

<table>
<thead>
<tr>
<th></th>
<th>#Tuple</th>
<th>#Rel Types</th>
<th>Node Type</th>
<th>Construction</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMCS</td>
<td>40K</td>
<td>21</td>
<td>Phrase &amp; Entity</td>
<td>Annotation</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>21M</td>
<td>36</td>
<td>Phrase &amp; Entity</td>
<td>Annotation+Auto</td>
</tr>
<tr>
<td>ATOMIC</td>
<td>880K</td>
<td>9</td>
<td>Free-text</td>
<td>Annotation</td>
</tr>
<tr>
<td>ATOMIC2020</td>
<td>1.33M</td>
<td>23</td>
<td>Free-text, Phrase &amp; Entity</td>
<td>Annotation</td>
</tr>
<tr>
<td>GLUCOSE</td>
<td>670K</td>
<td>10</td>
<td>Free-text, Structured Rules</td>
<td>Annotation</td>
</tr>
<tr>
<td>WebChild</td>
<td>4M</td>
<td>19</td>
<td>Phrase &amp; Entity</td>
<td>IR/IE</td>
</tr>
<tr>
<td>WebChild 2.0</td>
<td>18M</td>
<td>19</td>
<td>Phrase &amp; Entity</td>
<td>IR/IE</td>
</tr>
<tr>
<td>Quasimodo</td>
<td>2.26M</td>
<td>-</td>
<td>Phrase &amp; Entity</td>
<td>IR/IE</td>
</tr>
<tr>
<td>ASER (core)</td>
<td>52.3M</td>
<td>14</td>
<td>Eventuality (Activity, states, events)</td>
<td>IR/IE</td>
</tr>
<tr>
<td>TransOMCS</td>
<td>18.5M</td>
<td>20</td>
<td>Phrase &amp; Entity</td>
<td>IR/IE+Annotation+Reasoning</td>
</tr>
<tr>
<td>DISCOS</td>
<td>3.4M</td>
<td>9</td>
<td>Eventuality</td>
<td>IR/IE+Reasoning</td>
</tr>
</tbody>
</table>
Learning and Reasoning with CSKB/CSKG

• Introduction

• Learning and Reasoning on CSKBs/CSKGs
  • Commonsense Knowledge Bases
  • Commonsense Knowledge Generation
  • Commonsense Knowledge Base Completion
  • Commonsense Knowledge Base Population

• Learning and Reasoning for Downstream Tasks (CSQA)

Slides credit of this part: Tianqing Fang
Commonsense Generation

• Cloze style
  • LAMA
    • English ConceptNet, single-token objects.
    • \((Head, Relation, [MASK])\)

• Mining ConceptNet knowledge using PTLM
  • Turning triples to sentences
    • \((\text{ferret}, \text{AtLocation}, \text{pet store}) \rightarrow \text{ferret is in the pet store})\)
  • Generate tails using GPT and BERT

• A lot of prompt-based methods have been developed

Davison, Joe, Joshua Feldman, and Alexander M. Rush. “Commonsense knowledge mining from pretrained models.” EMNLP 2019
COMET: COMmonsEnse Transformers

- Train a transformer (GPT-2) of how to generate the tail
- Can be seen as a generative knowledge base population method
- How to generate/find new heads is unclear
Symbolic Knowledge Distillation

- Extracts the commonsense from the large, general language model GPT-3, into 2 forms:
  - a large commonsense knowledge graph ATOMIC\textsuperscript{10x}
  - a compact commonsense model COMET\textsuperscript{TIL}

**Prompt Heads**

1. Event: X overcomes evil with good
2. Event: X does not learn from Y
   ...
10. Event: X looks at flowers
11. ...

**Prompt Tails**

- A set of 100 high-quality events from ATOMIC\textsuperscript{20}
- Randomly sampling 10 each time
- Generate 165K unique events using the 175B-parameter Davinci model

For each pair of event (165K) and relation (7) we generate 10 inferences with the second largest form of GPT-3, Curie, resulting in 6.46M ATOMIC-style data triples.
Learning and Reasoning with CSKB/CSKG

• Introduction

• Learning and Reasoning on CSKBs/CSKGs
  • Commonsense Knowledge Bases
  • Commonsense Knowledge Generation
  • Commonsense Knowledge Base Completion
  • Commonsense Knowledge Base Population

• Learning and Reasoning for Downstream Tasks (CSQA)

Slides credit of this part: Tianqing Fang
Commonsense Knowledge Base Completion

• Commonsense Knowledge Base Completion
  • Adopt the idea of KB Completion
  • $\{(h, r, t)| h \in H, r \in R, t \in T\}$, predict missing links within the set of $H$ and $T$.

• Datasets:
  • ConceptNet
  • ATOMIC

• Differences with Conventional Knowledge Base Completion
  • Semantics matters a lot
  • Commonsense KBs are generally very sparse.
CSKB Completion

• CSKB Completion vs Traditional KB Completion

<table>
<thead>
<tr>
<th></th>
<th>#Nodes</th>
<th>#Edges</th>
<th>Avg In-Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConceptNet</td>
<td>78,088</td>
<td>10,000</td>
<td>1.25</td>
</tr>
<tr>
<td>ATOMIC</td>
<td>256,570</td>
<td>610,536</td>
<td>2.25</td>
</tr>
<tr>
<td>FB15K-237</td>
<td>14,505</td>
<td>272,115</td>
<td>16.98</td>
</tr>
</tbody>
</table>

• Need to deal with sparsity in CSKB.
• Need to encode semantics of the nodes.

CSKB Densification

• Bert-sim+GCN+Conv-TransE
• Graph densifier using BERT similarity
• GCN to encode graph structure
• Conv+a bilinear projection matrix decoder for link prediction

InductivE

- BERT+R-GCN+Conv-TransE (Modified)
  - R-GCN
  - Graph densifier using BERT similarity
  - Heuristic rules, adding edges for nodes with fewer neighbors

<table>
<thead>
<tr>
<th>Model</th>
<th>CN-100K</th>
<th>MRR</th>
<th>CN-82K</th>
<th>MRR</th>
<th>ATOMIC</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MRR</td>
<td>Hits@3</td>
<td>Hits@10</td>
<td></td>
<td>Hits@3</td>
</tr>
<tr>
<td>DistMult</td>
<td>10.62</td>
<td>10.94</td>
<td>22.54</td>
<td>2.80</td>
<td>2.90</td>
<td>5.60</td>
</tr>
<tr>
<td>ComplEx</td>
<td>11.52</td>
<td>12.40</td>
<td>20.31</td>
<td>2.60</td>
<td>2.70</td>
<td>5.00</td>
</tr>
<tr>
<td>ConvE</td>
<td>20.88</td>
<td>22.91</td>
<td>34.02</td>
<td>8.01</td>
<td>8.67</td>
<td>13.13</td>
</tr>
<tr>
<td>RotatE</td>
<td>24.72</td>
<td>28.20</td>
<td>45.41</td>
<td>5.71</td>
<td>6.00</td>
<td>11.02</td>
</tr>
<tr>
<td>COMET</td>
<td>6.07</td>
<td>2.92</td>
<td>21.17</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Malaviya et al.</td>
<td>52.25</td>
<td>58.46</td>
<td>73.50</td>
<td>16.26</td>
<td>17.95</td>
<td>27.51</td>
</tr>
<tr>
<td>InductivE</td>
<td>57.35</td>
<td>64.50</td>
<td>78.00</td>
<td>20.35</td>
<td>22.65</td>
<td>33.86</td>
</tr>
<tr>
<td>Improvement</td>
<td>9.8%</td>
<td>10.3%</td>
<td>6.1%</td>
<td>25.2%</td>
<td>26.2%</td>
<td>23.1%</td>
</tr>
</tbody>
</table>

**TABLE II:** Comparison of CKG completion results on CN-100K, CN-82K and ATOMIC datasets. Improvement is computed by comparing with [15].

Learning and Reasoning with CSKB/CSKG

• Introduction

• Learning and Reasoning on CSKBs/CSKGs
  • Commonsense Knowledge Bases
  • Commonsense Knowledge Generation
  • Commonsense Knowledge Base Completion
  • Commonsense Knowledge Base Population

• Learning and Reasoning for Downstream Tasks (CSQA)

Slides credit of this part: Tianqing Fang
CSKB Population

• Denote the CSKB as $\mathcal{C} = \{(h, r, t)|h \in \mathcal{H}, r \in \mathcal{R}, t \in \mathcal{T}\}$. An automatically extracted eventuality knowledge graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which is much larger than $\mathcal{C}$.

• Denote $\mathcal{G}^\mathcal{C}$ as the graph by aligning $\mathcal{G}$ and $\mathcal{C}$.

• The goal of CSKB Population is to learn a scoring function for a triple $(h, r, t)$ where plausible triples are scored higher.

• Triples from $\mathcal{C}$ are served as positive examples.
  • Graph propagation
  • Transductive learning
  • Linked to traditional semi-supervised learning as well
CKGC (Completion) vs. CKGP (Population)
Commonsense Knowledge Base Population

• Different commonsense knowledge bases have different properties

• ConceptNet Population
  • Selectional preference

• ATOMIC Population
  • Latent variables (events and states) of commonsense

Slides credit for this part: Hongming Zhang
ConceptNet (Speer & Havasi, 2012)
Core is OMCS (Liu & Singh 2004)

- Commonsense knowledge base
  - Commonsense knowledge about noun-phrases, or entities.

Revisit the Correlations of Selectional Preference and OMCS

- (sing, song) (dobj, 9.25)
- (song, UsedFor, sing)
- (phone, ring) (nsubj, 8.75)
- (phone, CapableOf, ring)
- (cold, water) (amod, 8.86)
- (water, HasProperty, cold)
- (create, new) (dobj_amod, 8.25)
- (create idea, UsedFor, invent new things)
- (hungry, eat) (nsubj_amod, 10.00)
- (eat, MotivatedByGoal, are hungry)
Revisit the Correlations of ASER and OMCS
TransOMCS

Relation: AtLocation
Pattern: \((H) \leftarrow \text{nsubj} \leftarrow ((T) \text{-obl} (at))\)
Knowledge: (Student, AtLocation, School)

Relation: Causes
Pattern: \((H) \leftarrow \text{dobj} \leftarrow () \leftarrow \text{Result} \leftarrow (T)\)
Knowledge: (Good grades, Causes, Graduate)

ASER Subgraph

- students
- school
- study
- case
- at
- obl
- he
- gets
- Result
- graduates
- nsubj
- dobj
- xcomp
- good
- colledge
- amod
Knowledge Ranking

• Assigning confidence score to each piece of extracted commonsense
  • Leverage the semantics of the original sentences
  • Leverage the frequency information
Transferring ASER to ConceptNet

<table>
<thead>
<tr>
<th>Model</th>
<th># Vocab</th>
<th># Tuple</th>
<th>Novel_t</th>
<th>Novel_c</th>
<th>ACC_n</th>
<th>ACC_o</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMET&lt;sub&gt;Original&lt;/sub&gt; (Greedy decoding)</td>
<td>715</td>
<td>1,200</td>
<td>33.96%</td>
<td>5.27%</td>
<td>58%</td>
<td>90%</td>
</tr>
<tr>
<td>COMET&lt;sub&gt;Original&lt;/sub&gt; (Beam search - 10 beams)</td>
<td>2,232</td>
<td>12,000</td>
<td>64.95%</td>
<td>27.15%</td>
<td>35%</td>
<td>44%</td>
</tr>
<tr>
<td>COMET&lt;sub&gt;Extended&lt;/sub&gt; (Greedy decoding)</td>
<td>3,912</td>
<td>24,000</td>
<td>99.98%</td>
<td>55.56%</td>
<td>34%</td>
<td>47%</td>
</tr>
<tr>
<td>COMET&lt;sub&gt;Extended&lt;/sub&gt; (Beam search - 10 beams)</td>
<td>8,108</td>
<td>240,000</td>
<td>99.98%</td>
<td>78.59%</td>
<td>23%</td>
<td>27%</td>
</tr>
<tr>
<td>LAMA&lt;sub&gt;Original&lt;/sub&gt; (Top 1)</td>
<td>328</td>
<td>1,200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>49%</td>
</tr>
<tr>
<td>LAMA&lt;sub&gt;Original&lt;/sub&gt; (Top 10)</td>
<td>1,649</td>
<td>12,000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>20%</td>
</tr>
<tr>
<td>LAMA&lt;sub&gt;Extended&lt;/sub&gt; (Top 1)</td>
<td>1,443</td>
<td>24,000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>29%</td>
</tr>
<tr>
<td>LAMA&lt;sub&gt;Extended&lt;/sub&gt; (Top 10)</td>
<td>5,465</td>
<td>240,000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>10%</td>
</tr>
<tr>
<td>TransOMCS&lt;sub&gt;Original&lt;/sub&gt; (no ranking)</td>
<td>33,238</td>
<td>533,449</td>
<td>99.53%</td>
<td>89.20%</td>
<td>72%</td>
<td>74%</td>
</tr>
<tr>
<td>TransOMCS (Top 1%)</td>
<td>37,517</td>
<td>184,816</td>
<td>95.71%</td>
<td>75.65%</td>
<td>86%</td>
<td>87%</td>
</tr>
<tr>
<td>TransOMCS (Top 10%)</td>
<td>56,411</td>
<td>1,848,160</td>
<td>99.55%</td>
<td>92.17%</td>
<td>69%</td>
<td>74%</td>
</tr>
<tr>
<td>TransOMCS (Top 30%)</td>
<td>68,438</td>
<td>5,544,482</td>
<td>99.83%</td>
<td>95.22%</td>
<td>67%</td>
<td>69%</td>
</tr>
<tr>
<td>TransOMCS (Top 50%)</td>
<td>83,823</td>
<td>9,240,803</td>
<td>99.89%</td>
<td>96.32%</td>
<td>60%</td>
<td>62%</td>
</tr>
<tr>
<td>TransOMCS (no ranking)</td>
<td>100,659</td>
<td>18,481,607</td>
<td>99.94%</td>
<td>&lt;u&gt;98.30%&lt;/u&gt;</td>
<td>54%</td>
<td>56%</td>
</tr>
<tr>
<td>OMCS in ConceptNet 5.0</td>
<td>36,954</td>
<td>207,427</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>&lt;u&gt;92%&lt;/u&gt;</td>
</tr>
</tbody>
</table>

Transferability from linguistic knowledge to commonsense knowledge

SP over eventualities can effectively represent interesting commonsense knowledge
Distribution of Relations and Accuracy

Distribution of Relations

Accuracy
Commonsense Knowledge Base Population

• ConceptNet Population
  • Selectional preference

• ATOMIC Population
  • Latent variables (events and states) of commonsense

Slides credit for this part: Tianqing Fang
Transform ASER to ATOMIC
Coverage and Implicit Edges

- Most event related commonsense relations are implicit on ASER
- ConceptNet (Event-related relations), ATOMIC, ATOMIC 2020, and GLUCOSE

<table>
<thead>
<tr>
<th></th>
<th>ASER_{norm} Coverage</th>
<th>Avg. Degree in ASER_{norm}</th>
<th>Avg. Degree in C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>head(%)</td>
<td>tail(%)</td>
<td>edge(%)</td>
</tr>
<tr>
<td>ATOMIC</td>
<td>79.76</td>
<td>77.11</td>
<td>59.32</td>
</tr>
<tr>
<td>ATOMIC_{20}</td>
<td>80.39</td>
<td>47.33</td>
<td>36.73</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>77.72</td>
<td>54.79</td>
<td>43.51</td>
</tr>
<tr>
<td>GLUCOSE</td>
<td>91.48</td>
<td>91.85</td>
<td>81.01</td>
</tr>
</tbody>
</table>

Table 3: The overall matching statistics for the four CSKBs. The edge column indicates the proportion of edges where their heads and tails can be connected by paths in ASER. Average (in and out)-degree on ASER_{norm} and C for nodes from the CSKBs is also presented. The statistics in C is different from (Malaviya et al., 2020) as we check the degree on the aligned CSKB C instead of each individual CSKB.
Node Alignment with ASER

- ASER and other CSKB take different forms of representing personal entities
- Develop simple rules for aligning the two resources.
DISCOS (DIScourse to COMmonSense): BertSAGE [WWW 2021]

- Use BERT to encode the eventuality sentences
- Use GraphSAGE (Hamilton 2017) to aggregate the neighboring information in ASER

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. DISCOS: Bridging the Gap between Discourse Knowledge and Commonsense Knowledge. WWW, 2021.
Another Model: KG-BertSAGE [EMNLP 2021]

Training and Testing Data

- Training: four commonsense knowledge bases
  - ConceptNet (event-related relations)
  - ATOMIC
  - ATOMIC 2020
  - GLUCOSE
- Graph Data: normalized nodes/edges in ASER
- Testing: ~30K annotated data
Main Population Results

- We use AUC as the evaluation metric. The break-down scores for all models are presented below.

<table>
<thead>
<tr>
<th>Relation</th>
<th>xWnt</th>
<th>oWnt</th>
<th>gWnt</th>
<th>xEfct</th>
<th>oEfct</th>
<th>gEfct</th>
<th>xRct</th>
<th>oRct</th>
<th>gRct</th>
<th>xAttr xInt</th>
<th>xNeed Cause</th>
<th>xRsn</th>
<th>isBfr</th>
<th>isAft</th>
<th>Hndr. HasSubE.</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>57.7</td>
<td>64.9</td>
<td>66.3</td>
<td>59.1</td>
<td>66.2</td>
<td>60.0</td>
<td>50.6</td>
<td>68.7</td>
<td>72.3</td>
<td>56.2</td>
<td>63.9</td>
<td>56.4</td>
<td>48.3</td>
<td>34.5</td>
<td>59.2</td>
<td>58.0</td>
</tr>
<tr>
<td>BERTSAGE</td>
<td>54.7</td>
<td>58.9</td>
<td>58.0</td>
<td>58.0</td>
<td>70.0</td>
<td>54.7</td>
<td>52.8</td>
<td>62.4</td>
<td>76.6</td>
<td>55.0</td>
<td>61.0</td>
<td>57.1</td>
<td>46.2</td>
<td>45.5</td>
<td>66.7</td>
<td>64.9</td>
</tr>
<tr>
<td>KG-BERT</td>
<td>63.2</td>
<td>69.8</td>
<td>69.0</td>
<td>68.0</td>
<td>70.6</td>
<td>61.0</td>
<td>57.0</td>
<td>64.0</td>
<td>73.8</td>
<td>59.5</td>
<td>64.9</td>
<td>64.6</td>
<td>47.4</td>
<td>90.9</td>
<td>78.0</td>
<td>77.5</td>
</tr>
<tr>
<td>KG-BERTSAGE</td>
<td>66.0</td>
<td>68.9</td>
<td>68.6</td>
<td>68.2</td>
<td>70.8</td>
<td>62.3</td>
<td>60.5</td>
<td>64.6</td>
<td>74.1</td>
<td>59.1</td>
<td>63.0</td>
<td>65.4</td>
<td>50.0</td>
<td>76.4</td>
<td>78.2</td>
<td>77.4</td>
</tr>
<tr>
<td>Human</td>
<td>86.2</td>
<td>86.8</td>
<td>83.3</td>
<td>85.2</td>
<td>83.9</td>
<td>79.8</td>
<td>81.1</td>
<td>82.6</td>
<td>76.5</td>
<td>82.6</td>
<td>85.6</td>
<td>87.4</td>
<td>80.1</td>
<td>73.7</td>
<td>89.8</td>
<td>89.9</td>
</tr>
</tbody>
</table>
GPT-2 (Generative) v.s. KG-Bert (Discriminative)

- Differences in the training setting. GPT-2: maximize the likelihood of positive examples. KG-Bert: distinguishing positive with (randomly sampled) negative examples. The former has better generalization ability.

<table>
<thead>
<tr>
<th>LR</th>
<th>all</th>
<th>Original Test Set</th>
<th>CSKB head + ASER tail</th>
<th>ASER edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>KGBert</td>
<td>67.5</td>
<td>79.2</td>
<td>57.3</td>
<td>52.6</td>
</tr>
<tr>
<td>KGBertSAGE</td>
<td>68.5</td>
<td>80.1</td>
<td>58.2</td>
<td>53.5</td>
</tr>
<tr>
<td>GPT2-small</td>
<td>70.5</td>
<td>73.3</td>
<td>64.0</td>
<td>63.0</td>
</tr>
<tr>
<td>GPT2-medium</td>
<td>71.5</td>
<td>74.7</td>
<td>65.1</td>
<td>65.1</td>
</tr>
<tr>
<td>GPT2-large</td>
<td>71.8</td>
<td>75.5</td>
<td>65.4</td>
<td>65.3</td>
</tr>
<tr>
<td>COMET(GPT2XL)</td>
<td>70.4</td>
<td>73.1</td>
<td>64.5</td>
<td>63.7</td>
</tr>
<tr>
<td>GPT2XL(ZS)</td>
<td>64.7</td>
<td>65.8</td>
<td>60.8</td>
<td>63.1</td>
</tr>
</tbody>
</table>
Learning and Reasoning with CSKB/CSKG

• Introduction

• Learning and Reasoning on CSKBs/CSKGs

• Learning and Reasoning for downstream tasks (CSQA)
  • Tasks and Resources for Commonsense Question Answering
  • Recent Methods for Commonsense Question Answering

Slides credit of this part: Zizheng Lin and Tianqing Fang
Overview

• Commonsense: the knowledge about the open world possessed by most people. (Liu and Singh, 2004)

• Example:
  • Amy gives the cellphone back to Bob after using it to call for her parents to pick her up.

Waiting for her parents ← Next action of Amy → Waiting for a new cellphone to be delivered

Much more likely than
Overview

• Commonsense Question Answering (CSQA):
  • Sophisticated comprehension
  • Complex reasoning

• CSQA Tasks and benchmarks:
  • Focus on one particular aspect (e.g., PIQA (Bisk et. al., 2020) for physical commonsense)
  • Covers general commonsense (e.g., CosmosQA (Huang et. al. 2020))

Overview

• Reporting bias: commonsense knowledge tends to be implicitly mentioned in unstructured data such as text

• CommonSense Knowledge Graphs (CSKG):
  • Provide explicit and structured commonsense knowledge
Tasks and Benchmarks

• Social commonsense
• Physical commonsense
• Temporal commonsense
• Numerical commonsense
• Spatial commonsense
• General commonsense
Social Commonsense

• Emotional and social intelligence required by human interactions in various social situations

• Example:
  • Alex spilled the food she just prepared all over the floor and it made a huge mess (Sap et al., 2019).

(a) Mop up the floor  ➔  Next action of Alex  ➔  (b) Taste the food  ➔  (c) Run around in the mess

Much more likely than
## Social Commonsense

<table>
<thead>
<tr>
<th>Sample Question</th>
<th>Sample Answer</th>
<th>Construction Method</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Social IQA</strong> (Sap et al., 2019)</td>
<td>In the school play, Robin played a hero in the struggle to the death with the angry villain. How would others feel afterwards?</td>
<td>ATOMIC, Human annotations</td>
<td>37.6K</td>
</tr>
<tr>
<td>(1) sorry for the villain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) hopeful that Robin will succeed ✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3) like Robin should lose</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SWAG</strong> (Zellers et al., 2018)</td>
<td>On stage, a woman takes a seat at the piano. She ___</td>
<td>ActivityNet Captions, Human annotation, Adversarial Filtering</td>
<td>113K</td>
</tr>
<tr>
<td>(1) sits on a bench as her sister plays with the doll</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) nervously sets her fingers on the keys ✓</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Physical Commonsense

• The common understanding of the physical properties of objects existing in everyday life

• Example:
  • The procedure of making an outdoor pillow (Bisk et al., 2020)

  blow into a trash bag and tie with rubber band
  blow into a tin can and tie with rubber band

Much more suitable than
# Physical Commonsense

<table>
<thead>
<tr>
<th>Sample Question</th>
<th>Sample Answer</th>
<th>Construction Method</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIQA (Bisk et al., 2020)</td>
<td>How do I find something I lost on the carpet?</td>
<td>(1) Put a solid seal on the end of your vacuum and turn it on. (2) Put a hair net on the end of your vacuum and turn it on. ✓</td>
<td>Instructions on everyday events</td>
</tr>
</tbody>
</table>
Temporal Commonsense

• Commonsense knowledge about time

• Example:

• taking a vacation

  takes longer time than

  taking a walk
## Temporal Commonsense

<table>
<thead>
<tr>
<th>Sample Question</th>
<th>Sample Answer</th>
<th>Construction Method</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCTACO (Zhou et al., 2019) Mr. Barco has refused US troops or advisors but has accepted US military aid. What happened after Mr. Barco accepted the military aid?</td>
<td>(1) the aid was denied (2) things started to progress ✓ (3) he received the aid ✓</td>
<td>Human annotations</td>
<td>13K</td>
</tr>
</tbody>
</table>

- Duration: how long an event takes
- Temporal ordering: typical order of events
- Frequency: how often an event occurs
- Stationarity: whether a state holds for a very long time or indefinitely

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. EMNLP/IJCNLP, 2019.
Numerical Commonsense

• Commonsense knowledge about numbers and their operations involved in everyday life.

• Example:
  • The number of days in a week
  seven
unnecessary to be explicitly mentioned in the communication
# Numerical Commonsense

<table>
<thead>
<tr>
<th>Sample Question</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NumerSense (Lin et al., 2020)</td>
<td>A bird usually has [MASK] legs.</td>
</tr>
<tr>
<td>DROP (Dua et al., 2019)</td>
<td>Before the UNPROFOR fully deployed, ..., and captured the village at 4:45 p.m. on 2 March 1992. The JNA ... the next day.</td>
</tr>
<tr>
<td></td>
<td>What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objects (35.2%)</td>
<td>A bicycle has <em>two</em> tires.</td>
</tr>
<tr>
<td>Biology (13.5%)</td>
<td>Ants have <em>six</em> legs.</td>
</tr>
<tr>
<td>Geometry (11.7%)</td>
<td>A cube has <em>six</em> faces.</td>
</tr>
<tr>
<td>Unit (6.3%)</td>
<td>There are <em>seven</em> days in a week.</td>
</tr>
<tr>
<td>Math (7.3%)</td>
<td>There are <em>nine</em> now.</td>
</tr>
<tr>
<td>Physics (5.7%)</td>
<td>There are <em>two</em> gases centigrade.</td>
</tr>
<tr>
<td>Geography (2.9%)</td>
<td>There are <em>seven</em> continents.</td>
</tr>
<tr>
<td>Misc. (17.5%)</td>
<td>There are <em>seven</em> states in the United States.</td>
</tr>
</tbody>
</table>

Table 1: NumerSense: Probing numerical commonsense knowledge of pre-trained language models.

- Subtraction
- Comparison
- Selection
- Addition
- Count
- Coreference
- Other arithmetic
- Etc.

• There are many other math word problems in NLP
Spatial Commonsense

• Cognitive process about spatial objects, relations, and transformations (Clements and Battista, 1992)

• Example:
  • The man is riding a horse (Collell et, al., 2018)

  The relative positions of the man and the horse

  The man is **above** the horse
### Spatial Commonsense

<table>
<thead>
<tr>
<th>Sample Question</th>
<th>Sample Answer</th>
<th>Construction Method</th>
<th>Size</th>
</tr>
</thead>
</table>
| **SPARTQA** *(Mirzaee et al., 2021)* | **STORY:** We have three blocks, A, B and C. Block B is to the right of block C and it is below block A. Block A has two black medium squares. Medium black square number one is below medium black square number two and a medium blue square. It is touching the bottom edge of this block. The medium blue square is below medium black square number two. Block B contains one medium black square. Block C contains one medium blue square and one medium black square. The medium blue square is below the medium black square.  
**QUESTIONS:**  
FB: Which block(s) has a medium thing that is below a black square? A, B, C  
FB: Which block(s) doesn't have any blue square that is to the left of a medium square? A, B  
FR: What is the relation between the medium black square which is in block C and the medium square that is below a medium black square that is touching the bottom edge of a block? Left  
CO: Which object is above a medium black square? the medium black square which is in block C or medium black square number two? medium black square number two  
YN: Is there a square that is below medium square number two above all medium black squares that are touching the bottom edge of a block? Yes | Human annotations and distant supervision | 140K |

---

General Commonsense

• General knowledge involved in everyday situation (e.g., causal commonsense)

• Example:

I tipped the bottle (Gordon et al., 2012)

The liquid in the bottle
poured out

What happened as a RESULT

The liquid in the bottle froze

Much more likely than

## General Commonsense

<table>
<thead>
<tr>
<th></th>
<th>Sample Question</th>
<th>Sample Answer</th>
<th>Construction Method</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>COPA (Gordon et, al., 2012)</td>
<td>The man fell unconscious. What was the cause of this?</td>
<td>(1) The assailant struck the man on the head. ✓</td>
<td>Human annotation</td>
<td>1k</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(2) The assailant took the man’s wallet.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CommonsenseQA (Talmor et, al., 2019)</td>
<td>Where can I stand on a river to see water falling without getting wet?</td>
<td>(1) waterfall, (2) bridge, ✓</td>
<td>Extraction from ConceptNet, Human annotation</td>
<td>12.2K</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(3) valley, (4) stream, (5) bottom</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CosmosQA (Huang et, al., 2019)</td>
<td>I cleaned xxx. His parents always throw our stuff like we were refugees. Why did I decide to clean?</td>
<td>(1) I’m getting tired (2) We gets more food and need rooms for that. ✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense re:
Learning and Reasoning with CSKB/CSKG

• Introduction

• Learning and Reasoning on CSKBs/CSKGs

• Learning and Reasoning for downstream tasks (CSQA)
  • Tasks and Resources for Commonsense Question Answering
  • Recent Methods for Commonsense Question Answering
    • Pre-Trained Language Model as the Only Implicit Knowledge Source
    • External Knowledge Graph as Explicit Knowledge Source
    • Induce Explicit Knowledge from Pre-Trained Language Model
    • Multitask Learning

Slides credit of this part: Zizheng Lin and Tianqing Fang
Pre-Trained Language Model as the Only Implicit Knowledge Source

• Pre-Trained Language Models (PTLMs) implicitly encode a certain amount of commonsense knowledge into its parameters by pre-training

• LAMA probe (Petroni et al., 2019):
  • Abundant knowledge can be induced from PTLMs via prompts
  • Inspired many following works studying the mechanism of inducing explicit knowledge from PTLMs

• Typical workflow:
  • Choose a PTLM (e.g., BERT, T5)
  • Formulate target questions into the chosen PTLM
  • Fine-tuning (Optional)
  • Prediction

Pre-Trained Language Model as the Only Implicit Knowledge Source

• UNICORN (Lourie et al., 2021)
  • T5-based CSQA model
  • Pre-trained and fine-tuned on a multi-task benchmark – RAINBOW (Lourie et al., 2021)
  • Sequential training paradigm
  • SOTA on various CSQA benchmarks (e.g., COSMOSQA and PIQA)
Learning and Reasoning with CSKB/CSKG

• Introduction

• Learning and Reasoning on CSKBs/CSKGs

• Learning and Reasoning for downstream tasks (CSQA)
  • Tasks and Resources for Commonsense Question Answering
  • Recent Methods for Commonsense Question Answering
    • Pre-Trained Language Model as the Only Implicit Knowledge Source
    • External Knowledge Graph as Explicit Knowledge Source
    • Induce Explicit Knowledge from Pre-Trained Language Model
    • Multi-task Learning

Slides credit of this part: Zizheng Lin and Tianqing Fang
External Knowledge Graph as Explicit Knowledge Source

• Reporting bias => PTLM alone may not be sufficient

• External knowledge graph => explicitly provide structured commonsense knowledge
**KagNet (Using ConceptNet)**

- 1. Concept Recognition from $Q$ and $A$.
- 2. Concept Matching in **ConceptNet**. Prepare a concept schema subgraph.
- 3. Path pruning using KG Embedding
- 4. GCN-LSTM-Attention

$Q$ for Questions and $A$ for Answers.
QA-GNN

- Scoring ConceptNet nodes with LMs to obtain the working graph
- Use Relational-GAT for working graph reasoning
ConceptNet+Wikipedia

• **XLNet + Graph Reasoning**
  
  1. Knowledge extraction (entity-based matching) from ConceptNet (less than 3 hops).
  2. Knowledge extraction (SRL) from Wikipedia. Using elastic search. \(<s, p>\) and \(<p, o>\) are added to the graph. \(s\) for subj, \(p\) for predicate, \(o\) for obj.
  3. Graph-Based Contextual Representation Learning. GCN + XLNet

DEKCOR (Using Wiktionary Descriptions)

- 1. Retrieve ConceptNet subgraph.
- 2. Extract context (description of entities) from Wiktionary.
- 3. Reasoning (Attention)

Casual Reasoning with Event Graph

• Using a Causal Event Graph (CEG) constructed from CausalBank Corpus
  • 314 million commonsense causal event pairs
• Retrieving related events to bridge implicit causations
• Using graph reasoning to perform inference
Learning and Reasoning with CSKB/CSKG

- Introduction

- Learning and Reasoning on CSKBs/CSKGs

- Learning and Reasoning for downstream tasks (CSQA)
  - Tasks and Resources for Commonsense Question Answering
  - Recent Methods for Commonsense Question Answering
    - Pre-Trained Language Model as the Only Implicit Knowledge Source
    - External Knowledge Graph as Explicit Knowledge Source
    - Induce Explicit Knowledge from Pre-Trained Language Model
    - Multitask Learning

Slides credit of this part: Zizheng Lin and Tianqing Fang
Induce Explicit Knowledge from Pre-Trained Language Model

• Self-Talk (Shwartz et al., 2020) paper pointed out LMs as knowledge providers suffer from various shortcomings:
  
  • **Insufficient coverage**: due to reporting bias, many trivial facts might not be captured by LMs because they are rarely written about
  
  • **Insufficient precision**: the distributional training objective increases the probability of non-facts that are semantically similar to true facts, as in negation (“birds cannot fly”)
  
  • **Limited reasoning capabilities**: it is unclear that LMs are capable of performing multiple reasoning steps involving implicit knowledge.
Unsupervised Commonsense Question Answering with Self-Talk

• 1. Generate a question, conditioned on the context (pink) and question prefix (yellow)
• 2. Generate an answer, conditioned on the context, generated question and a corresponding answer prefix
• 3. The clarification is a concatenation of the answer prefix and generated text (green).

WinoGrande Task
COMET-DynaGen (Bosselut et al., 2019)

- Inference in a zero-setting

Generate intermediate nodes with COMET

Evaluate each generated edge with conditional log-likelihood using COMET

Evaluate each answer edge with approximated PMI using COMET: removing the answer priors regardless of path (e.g., happy is a common answer to emotional reactions)
Learning and Reasoning with CSKB/CSKG

• Introduction

• Learning and Reasoning on CSKBs/CSKGs

• Learning and Reasoning for downstream tasks (CSQA)
  • Tasks and Resources for Commonsense Question Answering
  • Recent Methods for Commonsense Question Answering
    • Pre-Trained Language Model as the Only Implicit Knowledge Source
    • External Knowledge Graph as Explicit Knowledge Source
    • Induce Explicit Knowledge from Pre-Trained Language Model
    • Multitask Learning

Slides credit of this part: Zizheng Lin and Tianqing Fang
UnifiedQA

- **Text-to-text unification:**
  - Text in: [Question] + “\n” + ([Context], [Candidate Answers])
  - Text out: Answer

- Pre-trained on 8 QA datasets, SQuAD, NarrativeQA, RACE, ARC, etc.
  - Text-to-text PTLMs, BART and T5.
  - These pre-trained PTLM are then finetuned on each individual dataset for specific QAs.

---

UnifiedQA

• Text-to-text unification:
  • Performance of UnifiedQA (trained on all training set) and dedicatedly finetuned models on each individual dataset.
  • Performance v.s. directly finetuning PTLMs

<table>
<thead>
<tr>
<th></th>
<th>CommonsenseQA</th>
<th>WinoGrande</th>
<th>PIQA</th>
<th>SIQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART-FT</td>
<td>62.5</td>
<td>62.4</td>
<td>77.4</td>
<td>74.0</td>
</tr>
<tr>
<td>UnifiedQA-BART-FT</td>
<td>64.0</td>
<td>63.6</td>
<td>77.9</td>
<td>73.2</td>
</tr>
<tr>
<td>T5-FT</td>
<td>78.1</td>
<td>84.9</td>
<td>88.9</td>
<td>81.4</td>
</tr>
<tr>
<td>UnifeidQA-T5-FT</td>
<td>79.1</td>
<td>85.7</td>
<td>89.5</td>
<td>81.4</td>
</tr>
</tbody>
</table>
UNICORN

• 6 Multiple-choice based Commonsense QA datasets are merged.

• Training methods
  
  • Multi-task training: training on all multiple datasets (including the target dataset)
  
  • **Sequential training**: first training on multiple datasets (excluding the target dataset), and then continuing to train on the target dataset alone
  
  • Multi-task finetuning: first training on all datasets (including the target dataset), and then continuing to fine-tune on the target dataset alone

<table>
<thead>
<tr>
<th></th>
<th>αNLI</th>
<th>CosmosQA</th>
<th>HellaSWAG</th>
<th>PIQA</th>
<th>SIQA</th>
<th>WinoGrande</th>
</tr>
</thead>
<tbody>
<tr>
<td>multitask</td>
<td>78.4</td>
<td>81.1</td>
<td>81.3</td>
<td>80.7</td>
<td>74.8</td>
<td>72.1</td>
</tr>
<tr>
<td>finetune</td>
<td>79.2</td>
<td>82.6</td>
<td><strong>83.1</strong></td>
<td><strong>82.2</strong></td>
<td>75.2</td>
<td>78.2</td>
</tr>
<tr>
<td>sequential</td>
<td><strong>79.5</strong></td>
<td><strong>83.2</strong></td>
<td>83.0</td>
<td><strong>82.2</strong></td>
<td><strong>75.5</strong></td>
<td><strong>78.7</strong></td>
</tr>
<tr>
<td>none</td>
<td>77.8</td>
<td>81.9</td>
<td>82.2</td>
<td>80.2</td>
<td>73.8</td>
<td>77.0</td>
</tr>
</tbody>
</table>

UNICORN

• Due to reporting bias, commonsense rarely appears directly in text.
• Human annotated Commonsense Knowledge Bases (ConceptNet and ATOMIC) may provide additional info.
• Pretrain PTLM using constructing CSKBs.
• Task: Given \((h, r)\) predict \(t\), and given \((t, r)\) predict \(h\).

<table>
<thead>
<tr>
<th>CSKG</th>
<th>(\alpha)NLI</th>
<th>CosmosQA</th>
<th>HellaSWAG</th>
<th>PIQA</th>
<th>SIQA</th>
<th>WinoGrande</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOMIC</td>
<td>78.3</td>
<td>81.8</td>
<td>82.8</td>
<td>79.9</td>
<td>75.0</td>
<td>78.2</td>
</tr>
<tr>
<td>ConceptNet</td>
<td>78.0</td>
<td>81.8</td>
<td>82.5</td>
<td>80.5</td>
<td>74.3</td>
<td>76.3</td>
</tr>
<tr>
<td>Both</td>
<td>78.0</td>
<td>81.8</td>
<td>82.7</td>
<td>81.1</td>
<td>74.8</td>
<td>76.6</td>
</tr>
<tr>
<td>Single Task</td>
<td>77.8</td>
<td>81.9</td>
<td>82.8</td>
<td>80.2</td>
<td>73.8</td>
<td>77.0</td>
</tr>
</tbody>
</table>
Summary of Results

<table>
<thead>
<tr>
<th>Model</th>
<th>SWAG</th>
<th>SIQA</th>
<th>CosmosQA</th>
<th>PIQA</th>
<th>MCTACO</th>
<th>CommonsenseQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bert_{large}</td>
<td>86.6</td>
<td>64.5</td>
<td>66.8</td>
<td>66.7</td>
<td>42.72</td>
<td>56.7</td>
</tr>
<tr>
<td>XLNet_{large}</td>
<td>87.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RoBERTA_{large}</td>
<td>89.9</td>
<td>78.7</td>
<td>81.9</td>
<td>79.4</td>
<td>54.8</td>
<td>72.1</td>
</tr>
<tr>
<td>ALBERT_{XXL}</td>
<td>90.7</td>
<td>-</td>
<td>85.4</td>
<td>-</td>
<td>-</td>
<td>83.3</td>
</tr>
<tr>
<td>T5_{11B}</td>
<td>-</td>
<td>81.4</td>
<td>90.3</td>
<td>88.9</td>
<td>-</td>
<td>78.1</td>
</tr>
<tr>
<td>UnifiedQA</td>
<td>-</td>
<td>81.5</td>
<td>-</td>
<td>89.5</td>
<td>-</td>
<td>79.1</td>
</tr>
<tr>
<td>UNICORN</td>
<td>-</td>
<td>83.2</td>
<td>91.8</td>
<td>90.1</td>
<td>-</td>
<td>79.3</td>
</tr>
<tr>
<td>BERT-large</td>
<td>86.6</td>
<td>64.5</td>
<td>66.8</td>
<td>66.7</td>
<td>42.72</td>
<td>56.7</td>
</tr>
<tr>
<td>XLNet-large</td>
<td>87.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RoBERTA-large</td>
<td>89.9</td>
<td>78.7</td>
<td>81.9</td>
<td>79.4</td>
<td>54.8</td>
<td>72.1</td>
</tr>
<tr>
<td>ALBERT_{XXL}</td>
<td>90.7</td>
<td>-</td>
<td>85.4</td>
<td>-</td>
<td>-</td>
<td>83.3</td>
</tr>
<tr>
<td>T5_{11B}</td>
<td>-</td>
<td>81.4</td>
<td>90.3</td>
<td>88.9</td>
<td>-</td>
<td>78.1</td>
</tr>
<tr>
<td>UnifiedQA</td>
<td>-</td>
<td>81.5</td>
<td>-</td>
<td>89.5</td>
<td>-</td>
<td>79.1</td>
</tr>
<tr>
<td>UNICORN</td>
<td>-</td>
<td>83.2</td>
<td>91.8</td>
<td>90.1</td>
<td>-</td>
<td>79.3</td>
</tr>
</tbody>
</table>

- PTLMs achieve SOTA performance on five out of the six benchmarks
- PTLMs store a sufficient amount of commonsense knowledge for many CSQA tasks
- Pre-training giant models on large-scale corpora indeed benefits many CSQA tasks
- May not be sufficient for temporal CSQA yet
- UNICORN is the most competitive among PTLMs
- Multi-task training followed by task-specific fine-tuning paradigm
- Self-Talk model can improve zero-shot learning
- BERT-large model has very low scores on several datasets:
  - Under-trained issue

COMET-DynaGen: - 52.6 - - - -
Timeline of Approaches

Bi-Linear KG-Embedding
Li et al, 2016, Saito et al. 2018, Jastrzebski et al. 2018

KagNet
Lin, et al 2019

HyKAS 2.0
Ma, et al 2019

XLNet+
Reason
Lv et al 2020

Bert-similarity+
GCN+
Conv-TransE
Malaviya, et al, 2020

QA-GNN
Yasunaga, et al 2020

TransOMCS
Zhang, et al, 2020

DISCOS
Fang, et al, 2021

Neuro-Symbolic KG Completion
Moghimifar, et al, 2021

UNICORN
Lourie, et al, 2021

BENCHMARKING
Fang, et al, 2021

CSKB Population

CSKB Completion

Knowledge-enhanced

Multi-task

PTLM

2018 and before

2019

2020

2021

UnifiedQA
Khashabi, et al, 2021

DISCOS
Fang, et al, 2021

BENCHMARKING
Fang, et al, 2021

CSKB Population

CSKB Completion

Knowledge-enhanced

Multi-task

PTLM
Abductive Natural Language Inference

• Deductive reasoning and abductive reasoning thus differ in which end, left or right, of the proposition “X entails Y” serves as conclusion.
  • Deduction: from X to Y: e.g., All sharks have teeth, Alice is a shark → Alice has teeth
  • Abduction: from Y to find a set of explanations X that is consistent with some logical theory Z

\[
\alpha_{NLI}/\alpha_{NLG}\text{Data}
\]

O1: The observation at time t1
O2: The observation at time t2 > t1
h+: A plausible hypothesis that explains the two observations O1 and O2
h −: An implausible (or less plausible) hypothesis for observations O1 and O2

\[
h^* = \arg \max_{h^i} P(H = h^i | O_1, O_2)
\]

Difference between linear chain and fully connected model:
O1: “Carl went to the store desperately searching for flour tortillas for a recipe.”
O2: “Carl left the store very frustrated.”
h1: “The cashier was rude” (linear chain choose this) incorrect
h2: “The store had corn tortillas, but not flour ones.” (fully connected choose this) correct

Commonsense Inference of Dialogues

- Annotated 19 ConceptNet relations (e.g., Capable Of, Causes, Motivated By Goal) and 6 negated relations (Not Causes, Not Motivated By Goal)
- 807 dialogues from Daily Dialog, MuTual, DREAM
  - 5-12 utterances in each dialogue
- Several tasks: Dialogue-level Natural Language Inference, Span Extraction, Multi-choice Span Selection
Visual Commonsense Knowledge Graphs

- Sink in the water.
- Try to help [Person2].
- Save himself from drowning.
- Wait for help to arrive.
- Notice water washing in.
- Swim towards the statute.
- Sense his own death.
- Be washed away.
- Scream for help.
- Gasp for air.
- Before Person2 needed to ...
- After Person2 will most likely ...
- Befor Person2 needed to ...  
- Before Person1 needed to ...
- After Person1 will most likely ...
- Because Person1 wanted to ...
- Because Person2 wanted to ...
- Get to the top of the deck.
- Realize the ship is sinking.
- Start moving against the water.
- Get caught in a rush of water.

Park et al., 2020
Conclusions and Future Works

• Commonsense acquisition: we now have larger scale of
  • Crowdsourcing
  • Information extraction from the Web

• Large language models have been proven to be powerful for induction in a domain defined and designed by human
  • Even it’s open domain
  • The patterns that crowdsourcing workers annotate are supervised by the data creator
  • But we don’t know yet how to perform explicit reasoning on modern datasets/tasks

• Fundamentally, we regard following things are important for the future of developing commonsense reasoning
  • Conceptualization/abstraction: probabilistic modeling
  • Partial information aggregation and typicality inference
  • Larger commonsense evaluation datasets
    • Especially those cannot be solved by giant language models
  • Theory of mind mapped to practical computation
The Future of Commonsense Reasoning: Many are still missing!

Thank you for your attention! 😊