香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Commonsense Knowledge Acquisition and Reasoning

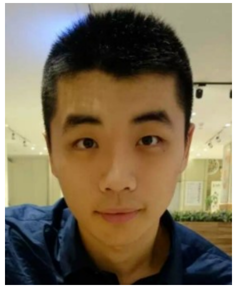
Yangqiu Song

Department of CSE, HKUST

Special thanks to

Tianqing Fang

Zizheng Lin



Hongming Zhang

for their contribution of slides.

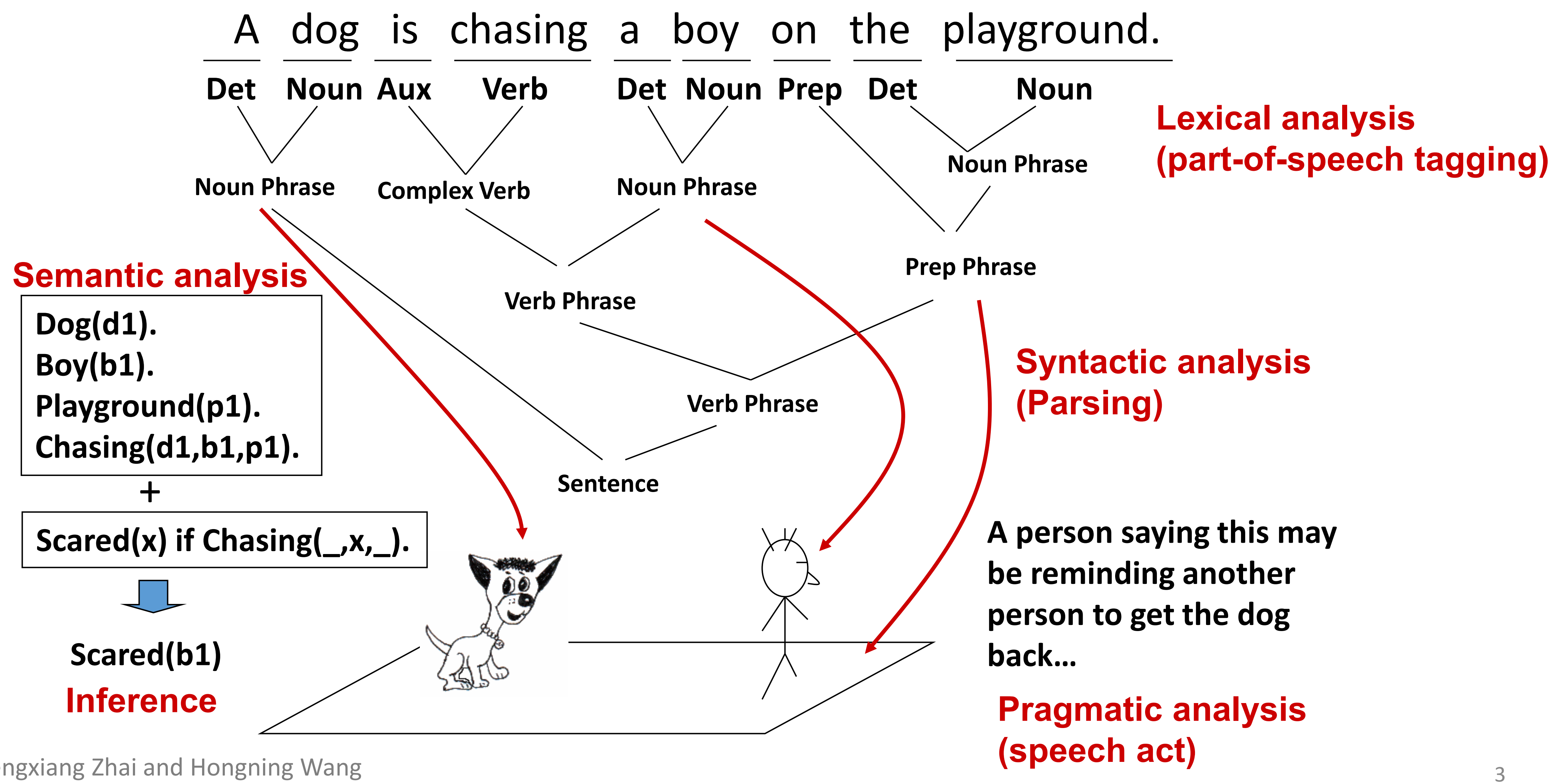
Understanding Human's Language Requires Complex Knowledge

- "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)

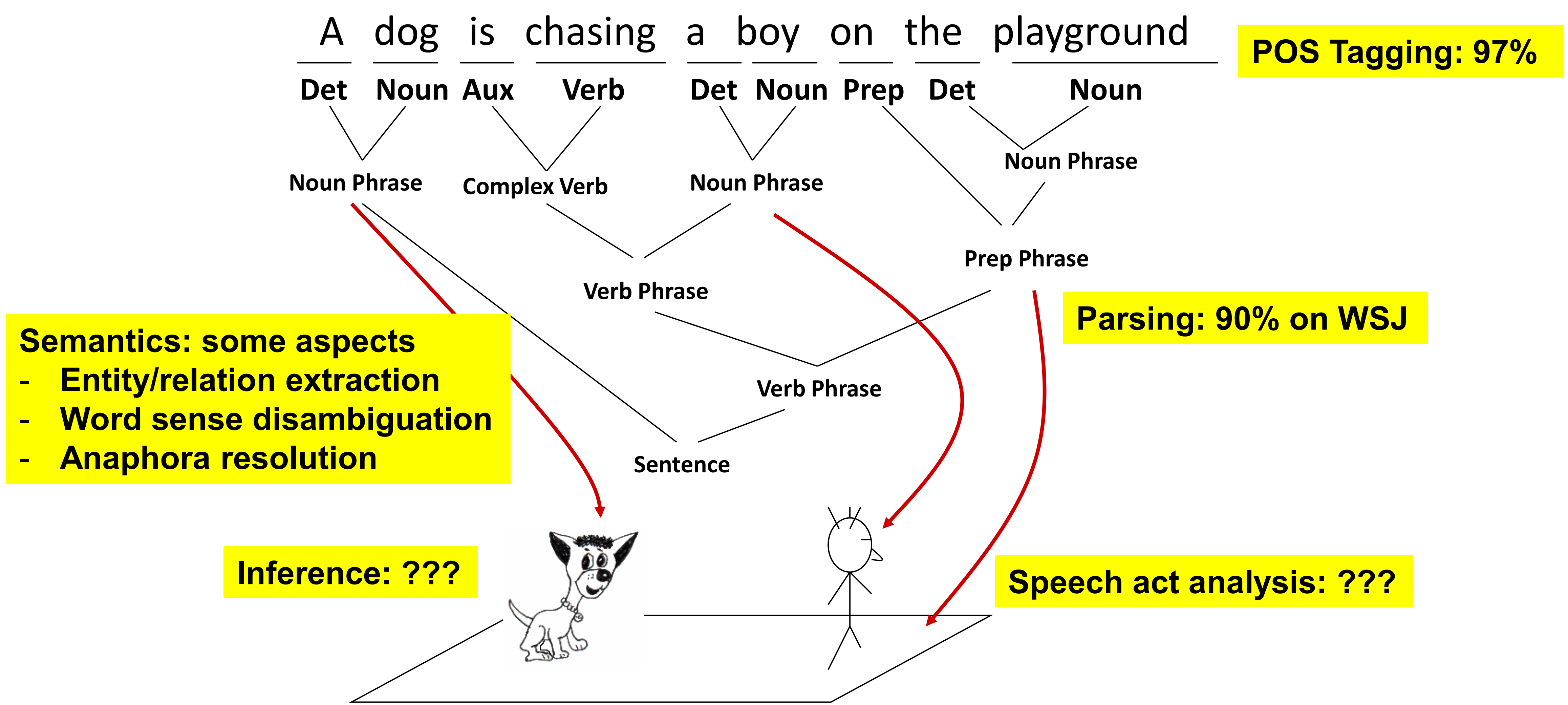
- Contexts and knowledge contributes to the meanings

https://www.thoughtco.com/world-knowledge-language-studies-1692508

An Example of NLP



The State of the Art



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.

- There is someone/something in danger.
- They are cooperating to save (the case).

- 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.

“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', "

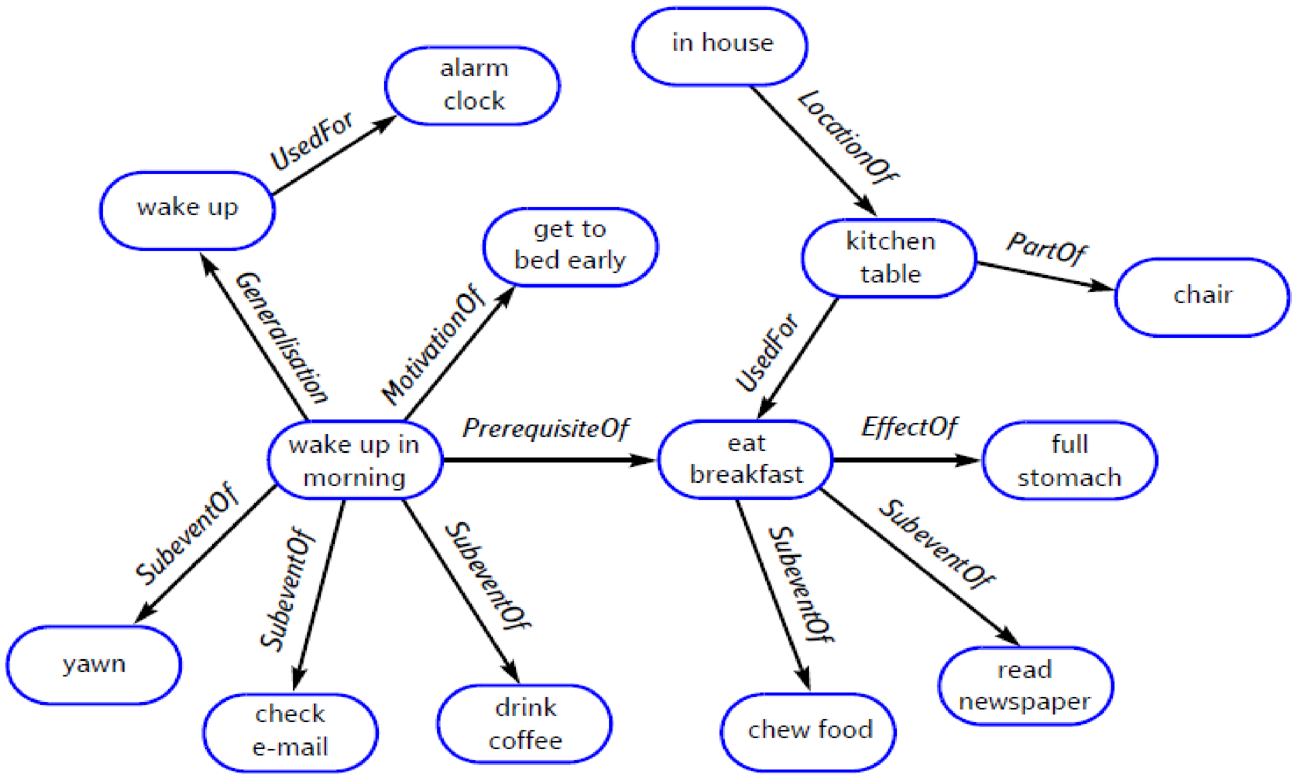
- "the AI community it is used in a technical sense to refer to the millions of basic facts and understandings possessed by most people."

- "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

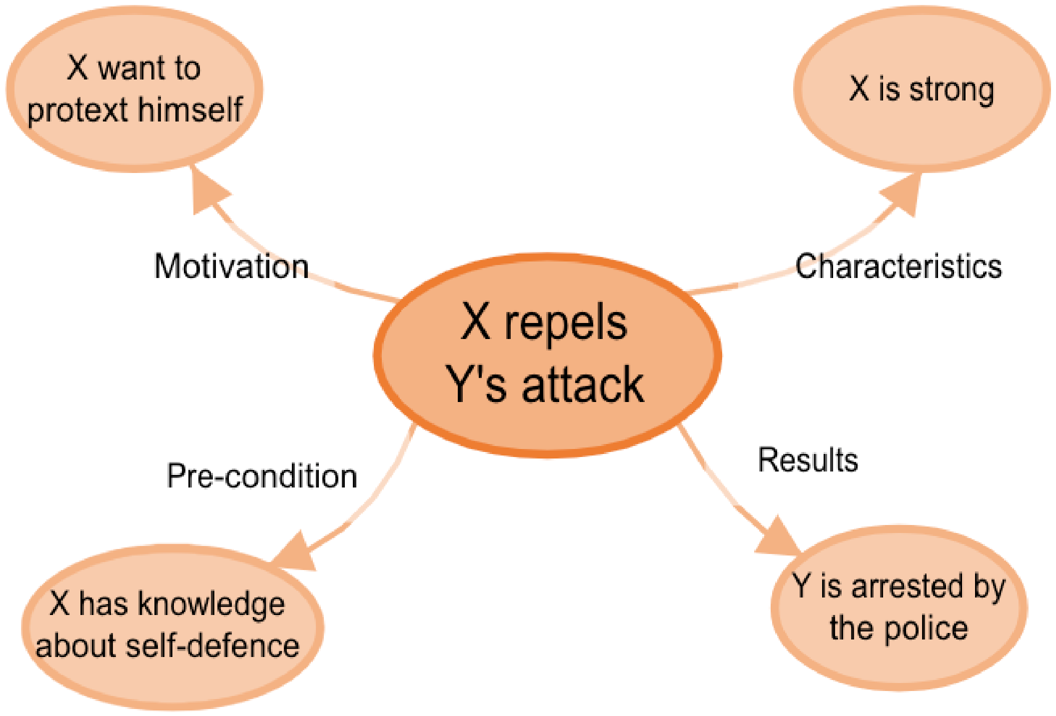
- ConceptNet5 (Speer and Havasi, 2012)
  - Core is from Open Mind Common Sense (OMCS) (Liu & Singh, 2004)



Essentially a crowdsourcing based approach + text mining

ATOMIC: Everyday If-then Commonsense Knowledge

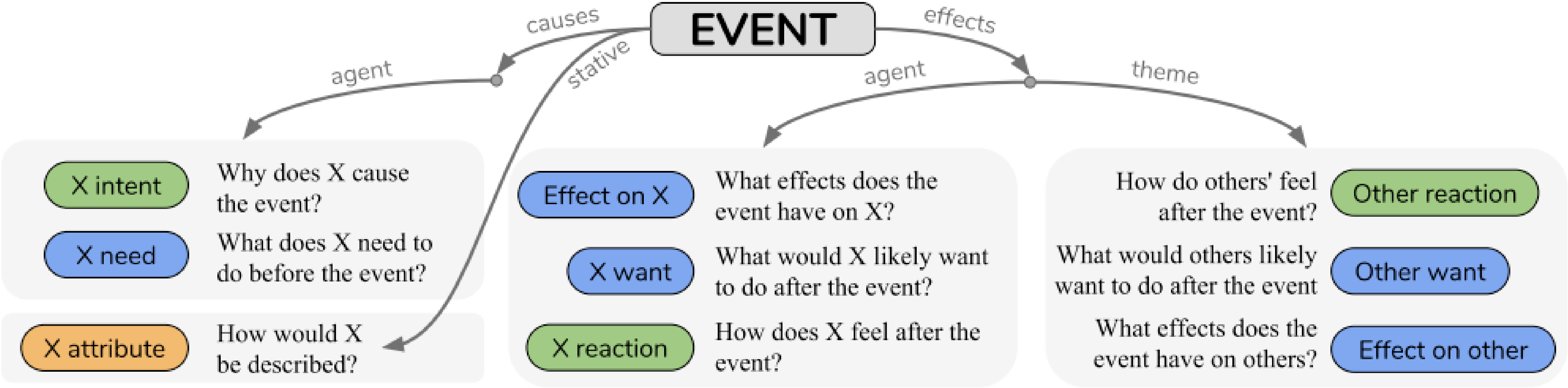
- These are day-to-day knowledge that help us understand each other.
- If a person *X did* something, human beings are able to inference:
  - Motivation: Why person X did this.
  - Pre-conditions: What enables X to do this.
  - Characteristics: What are attributes of X.
  - Result: What will affect X/others



Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, Yejin Choi: ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. AAAI, 2019.

ATOMIC: Everyday If-then Commonsense Knowledge

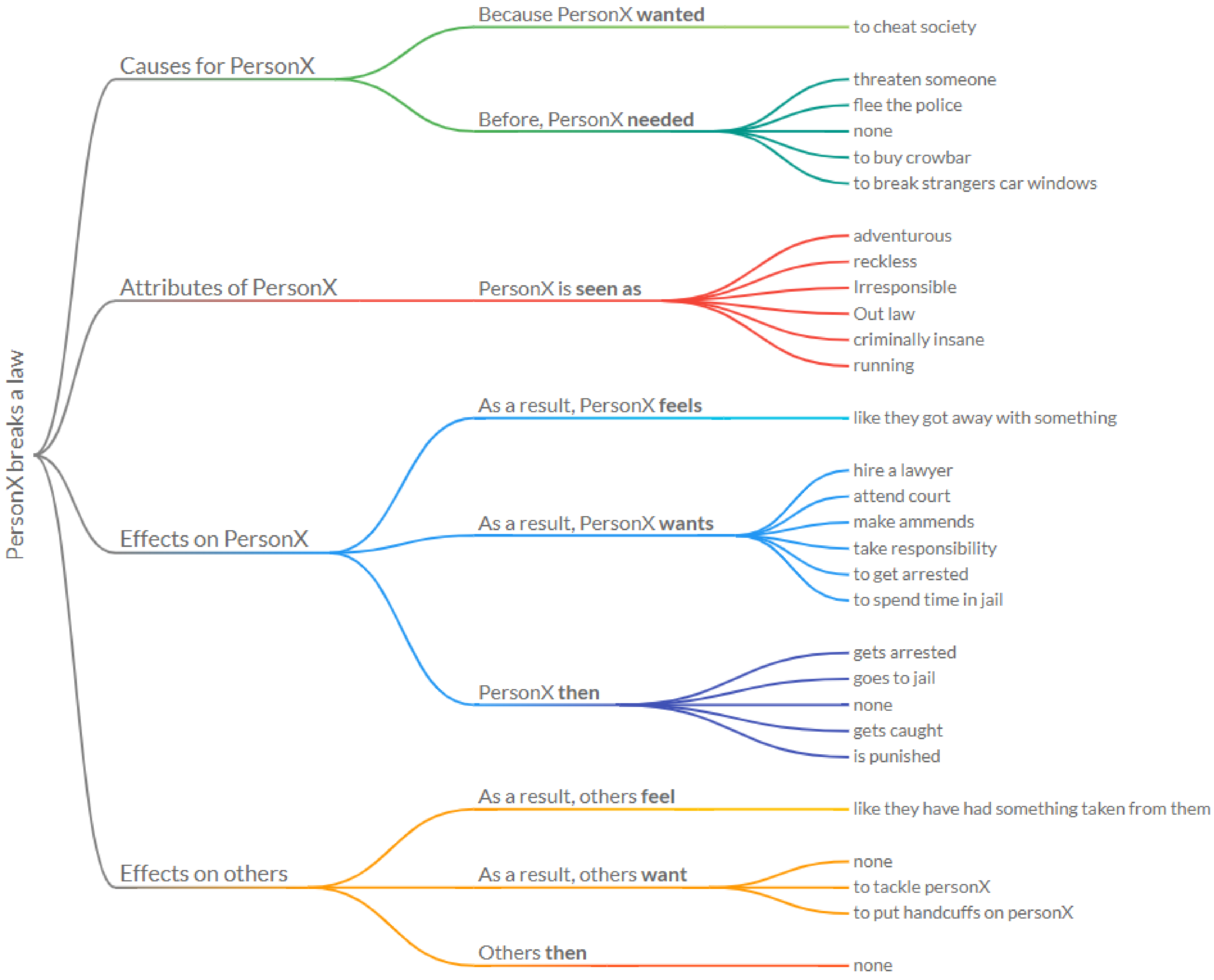
- Define 4 categories of if-then relations:
  - Causes-agent (Motivation & Pre-condition): xIntend, xNeed
  - Stative (Characteristics): xAttr
  - Effects-agent (Results on X): xWant, xReact, xEffect
  - Effects-theme (Results on others): oWant, oReact, oEffect



Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, Yejin Choi: ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. AAAI, 2019.

ATOMIC

- Crowdsourcing 9 Types of IF-THEN relations
- Arbitrary texts: Human annotation
- All personal entity information has been removed to reduce ambiguity



Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, Yejin Choi: ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. AAAI, 2019.

Ways of Collecting Commonsense Knowledge

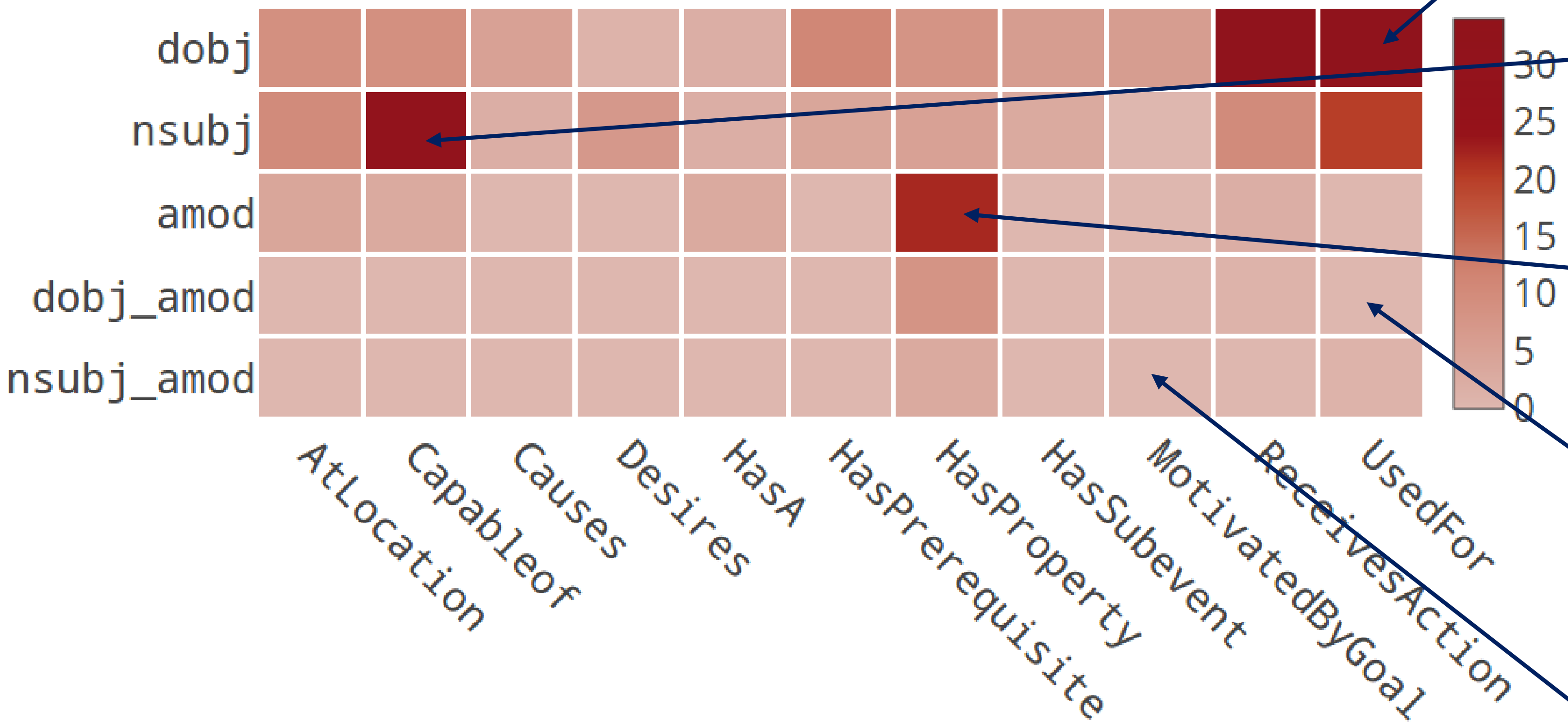
- Crowdsourcing
  - Pros
    - High quality
      - With proper quality control
    - Human can be creative when writing answers
      - Reflecting the ambiguity of language use
  - Cons
    - Ways of collection will limit the objects
      - Training Turk users: overfitting to the supervisor?
      - 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 population 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)

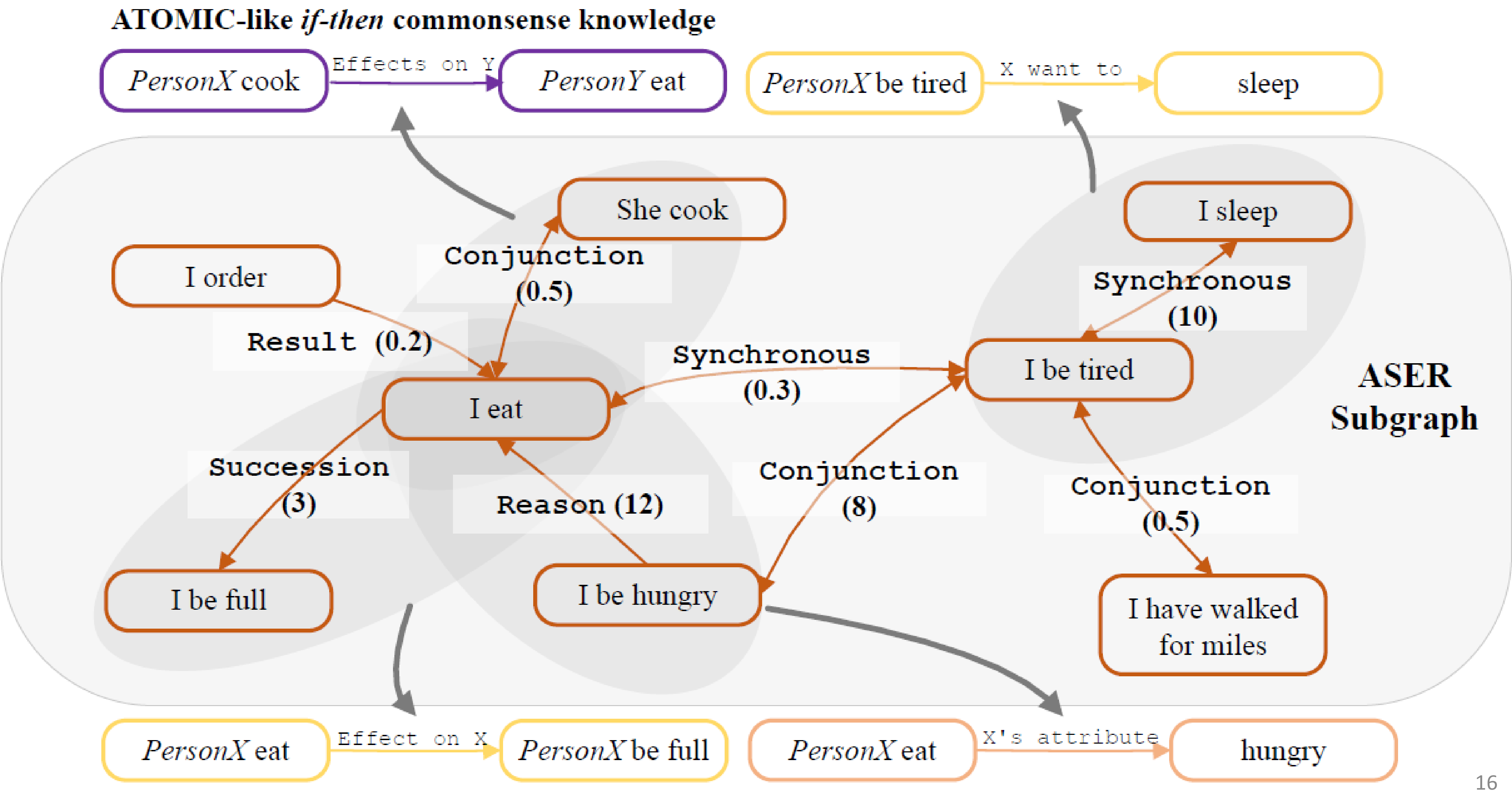
(phone, ring) (nsubj, 8.75)
(phone, CapableOf, ring)

(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

	ASER_{norm} Coverage				Avg. Degree in ASER_{norm}				Avg. Degree in $\mathcal{C}$			
					In-Degree		Out-Degree		In-Degree		Out-Degree	
	head(%)	tail(%)	edge(%)	#hops	head	tail	head	tail	head	tail	head	tail
ATOMIC	79.76	77.11	59.32	2.57	90.9	61.3	91.2	61.6	4.2	3.4	34.6	1.5
ATOMIC_{20}^{20}	80.39	47.33	36.73	2.65	96.9	66.9	97.3	67.3	4.3	2.9	34.6	1.5
ConceptNet	77.72	54.79	43.51	2.37	210.7	88.9	211.6	88.9	15.1	8.0	26.2	4.1
GLUCOSE	91.48	91.85	81.01	2.37	224.9	246.4	226.6	248.0	7.2	7.7	6.7	5.5

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 $\mathcal{C}$ for nodes from the CSKBs is also presented. The statistics in $\mathcal{C}$ is different from (Malaviya et al., 2020) as we check the degree on the aligned CSKB $\mathcal{C}$ instead of each individual CSKB.

Maarten Sap, et al. ATOMIC: An atlas of machine commonsense for if-then reasoning. AAAI 2019.
Jena D Hwang, et al. (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. AAAI 2021.
Nasrin Mostafazadeh, et al. Glucose: Generalized and contextualized story explanations. NAACL 2020.

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

Share

Mask 1

Prediction	Score
If you forget someone ’ s birthday , they may be **angry** with you .	40.2%
If you forget someone ’ s birthday , they may be **upset** with you .	10.6%
If you forget someone ’ s birthday , they may be **furious** with you .	8.3%
If you forget someone ’ s birthday , they may be **disappointed** with you .	7.1%
If you forget someone ’ s birthday , they may be **annoyed** with you .	2.9%

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

GPT-2

Sentence

If you forget someone's birthday,

Run Model

Model Output

Share

Prediction	Score
If you forget someone's birthday, **you can tell them it ...**	98.1%
If you forget someone's birthday, **or you're confused or ...**	1.4%
If you forget someone's birthday, **let's change it for ...**	0.5%
If you forget someone's birthday, **the customer will be left ...**	0%
If you forget someone's birthday, **the cheque is not ...**	0%

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

BERT

Sentence

To open a door, you must usually first turn the [MASK].

Run Model

Model Output

Share

Mask 1

Prediction	Score
To open a door , you must usually first turn the **knob** .	69.6%
To open a door , you must usually first turn the **key** .	11.9%
To open a door , you must usually first turn the **lock** .	9.9%
To open a door , you must usually first turn the **handle** .	7.3%
To open a door , you must usually first turn the **locks** .	0.5%

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

GPT-2

Sentence

To open a door, you must usually first

Run Model

Model Output

Share

Prediction	Score
To open a door, you must usually first **go to the door.**	97%
To open a door, you must usually first **listen for the sounds of ...**	2.7%
To open a door, you must usually first **void the door with the ...**	0.3%
To open a door, you must usually first **square the room with your ...**	0%
To open a door, you must usually first **connect the pipes and doors ...**	0%

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

GPT-2

Sentence

To open a door, you must usually first turn

Run Model

Model Output

Share

Prediction	Score
To open a door, you must usually first turn **your head so that you ...**	64.6%
To open a door, you must usually first turn **around and walk away.**	34%
To open a door, you must usually first turn **to the left, through ...**	0.9%
To open a door, you must usually first turn **around to get a grip ...**	0.3%
To open a door, you must usually first turn **the small door open and ...**	0.2%

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

BERT

Sentence

A lemon is [MASK].

Run Model

Model Output

Share

Mask 1

Prediction	Score
A lemon is **used** .	18.9%
A lemon is **eaten** .	4.9%
A lemon is **common** .	4%
A lemon is **preferred** .	3.4%
A lemon is **edible** .	1.8%

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

BERT

Sentence

Lemon is [MASK].

Run Model

Model Output

Share

Mask 1

Prediction	Score
Lemon is **used** .	7.7%
Lemon is **eaten** .	6.3%
Lemon is **preferred** .	6.3%
Lemon is **common** .	4.4%
Lemon is **added** .	2.4%

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

BERT

Sentence

A lemon is a [MASK].

Run Model

Model Output

Share

Mask 1

Prediction	Score
A lemon is a **lemon** .	9.9%
A lemon is a **fruit** .	5.8%
A lemon is a **candy** .	5%
A lemon is a **dessert** .	3.4%
A lemon is a **plant** .	2.4%

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

BERT

Sentence

Lemon is a [MASK].

Run Model

Model Output

Share

Mask 1

Prediction	Score
Lemon is a **nickname** .	1.6%
Lemon is a **synonym** .	1.2%
Lemon is a **surname** .	1.2%
Lemon is a **verb** .	1%
Lemon is a **pseudonym** .	0.9%

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

BERT

Sentence

The taste of lemon is [MASK].

Run Model

Model Output

Share

Mask 1

Prediction	Score
The taste of lemon is **sweet** .	30.4%
The taste of lemon is **bitter** .	17.4%
The taste of lemon is **distinctive** .	5.2%
The taste of lemon is **unpleasant** .	3.7%
The taste of lemon is **pleasant** .	2.9%

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

So Far We Know That

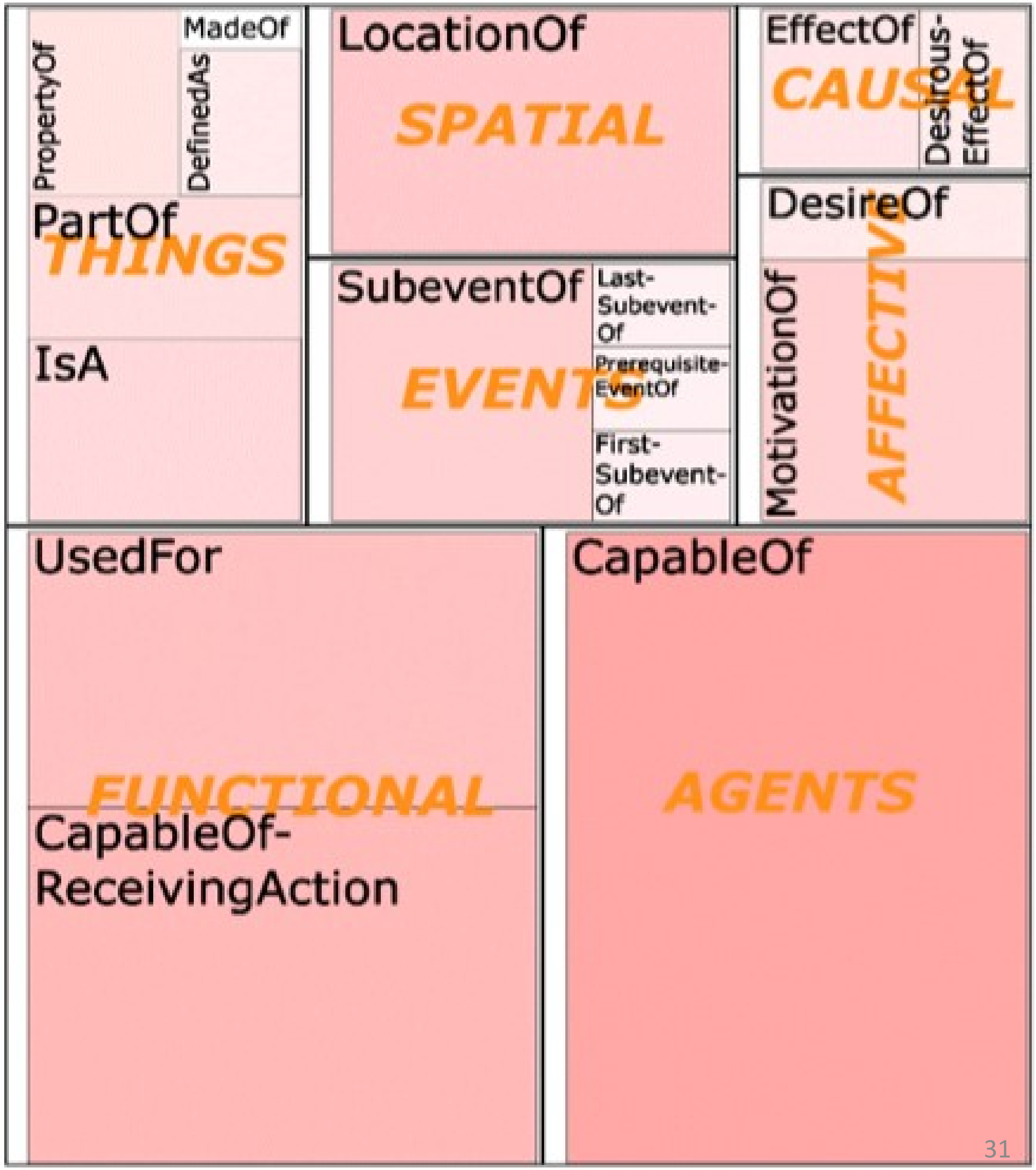
- Some commonsense may appear in selectional preference when we talk
- Event and casual relations: explicit extraction may not be useful for commonsense
    - More inference and/or reasoning have to be performed
- Large languages models probably need appropriate use (prompt) to get commonsense knowledge

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

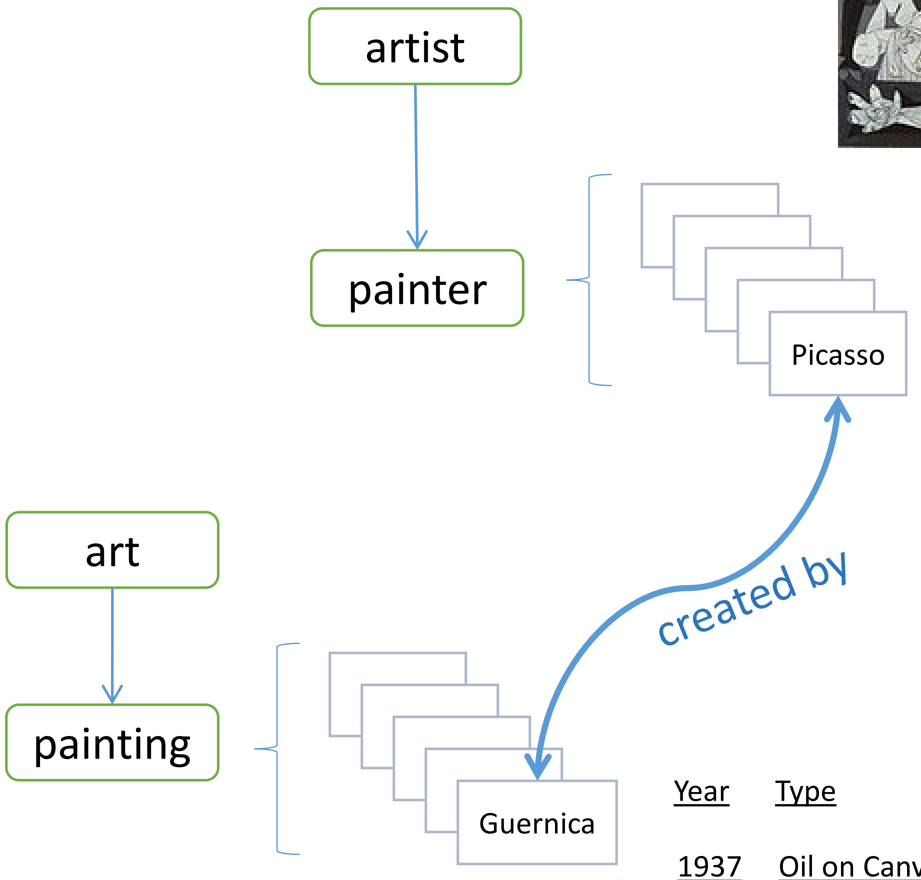


Primitive Semantic Units in our Mind

- 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.

Jackendoff, R. (Ed.). (1990). Semantic Structures. Cambridge, Massachusetts: MIT Press.

Knowledge Base



Born Died ... Movement

1881 1973 ... Cubism

Year Type ...

1937 Oil on Canvas ...

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

Google Knowledge Graph (2012)
570 million entities and 18 billion facts

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

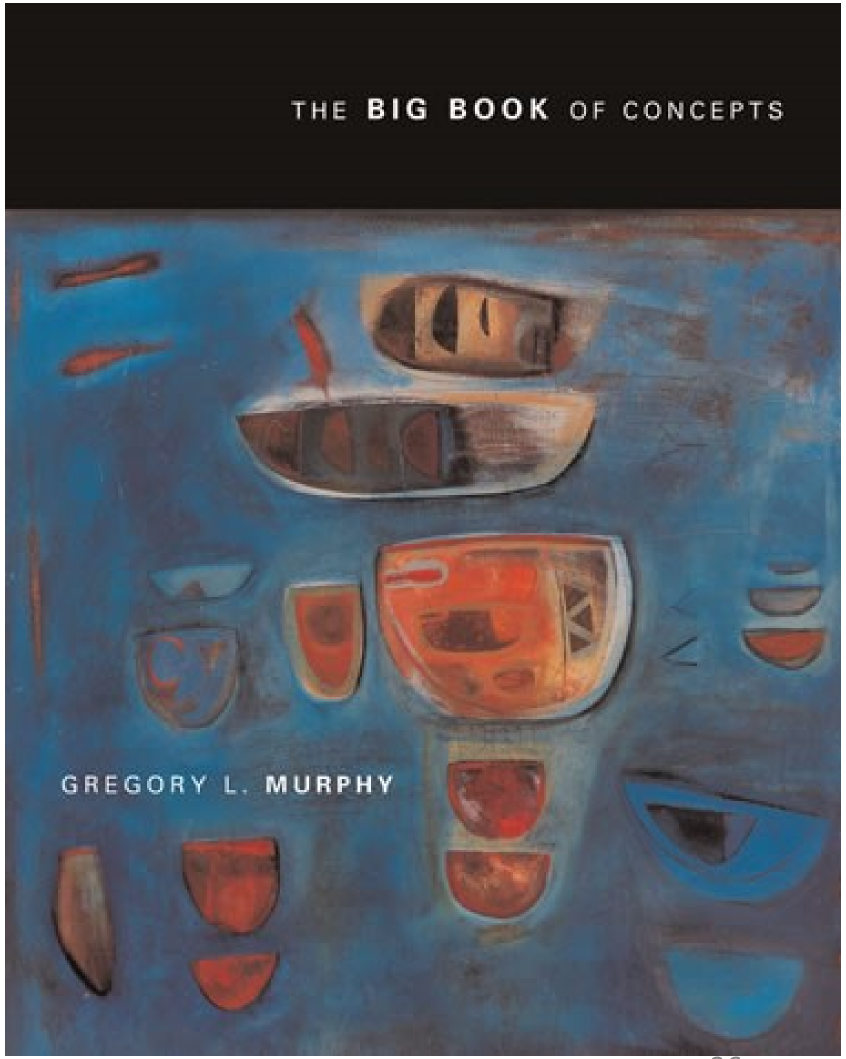
How to grow a mind: statistics, structure, and abstraction. Science. Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, Noah D Goodman. 2011.

“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”

bird



Slide Credit: Haixun Wang

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

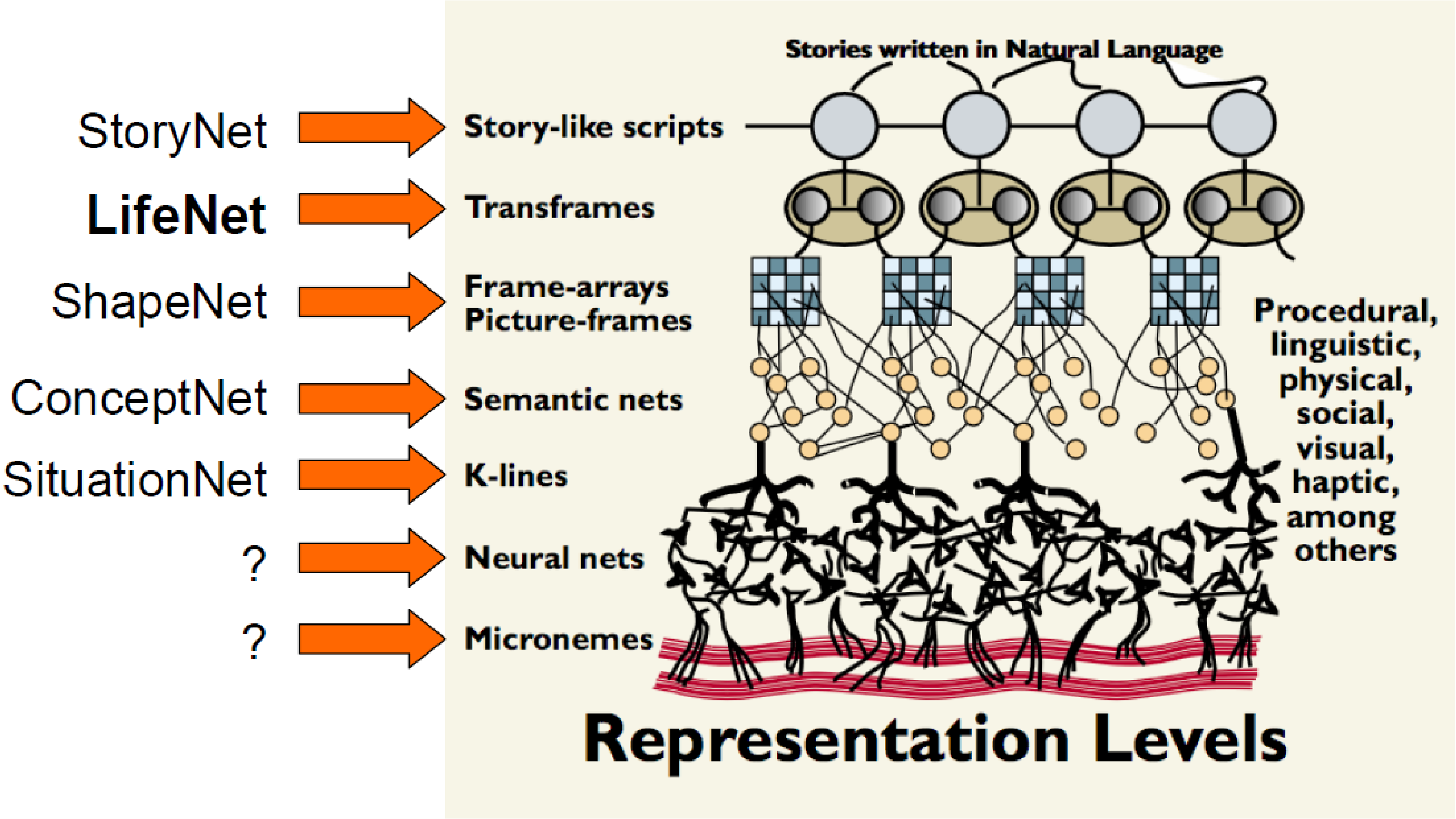
Our projects

- 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)

Push Singh 16/22 MIT Media Lab

https://ocw.mit.edu/courses/media-arts-and-sciences/mas-961-ambient-intelligence-spring-2005/lecture-notes/week4_push_singh.pdf

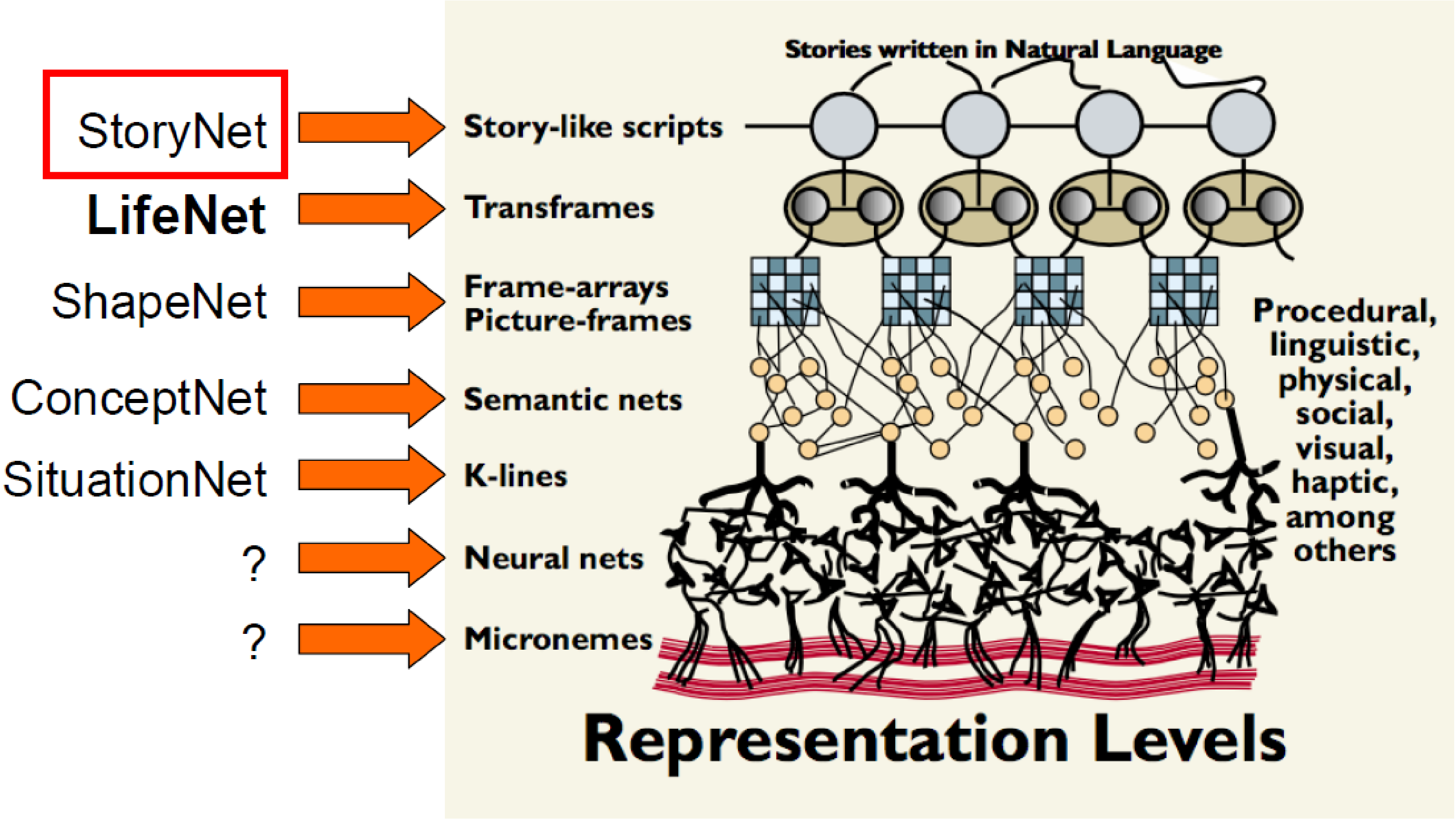
Representing Knowledge in Multiple Ways



Singh, Barbara Barry, and Hugo Liu (2004). **Teaching machines about everyday life.** *BT Technology Journal*, 22(4):227-240.

Representing Knowledge in Multiple Ways

StoryNet



Singh, Barbara Barry, and Hugo Liu (2004). **Teaching machines about everyday life.** *BT Technology Journal*, 22(4):227-240.

Representing Knowledge in Multiple Ways

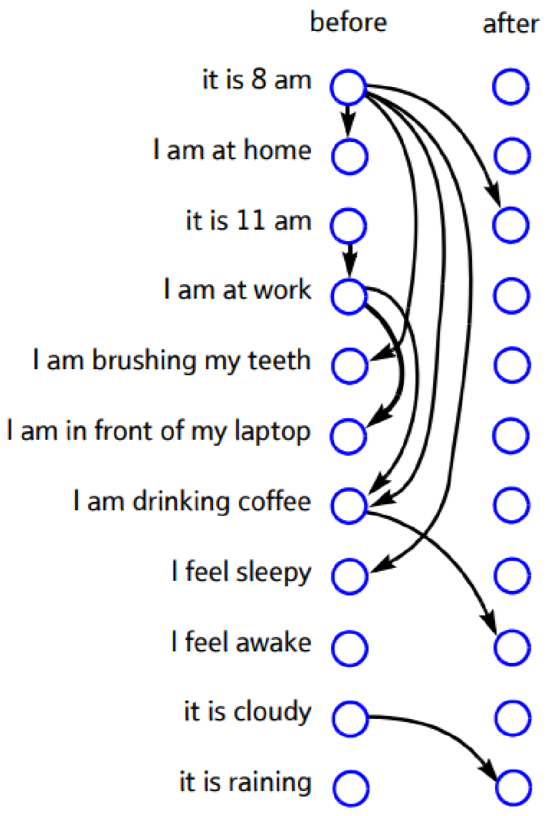
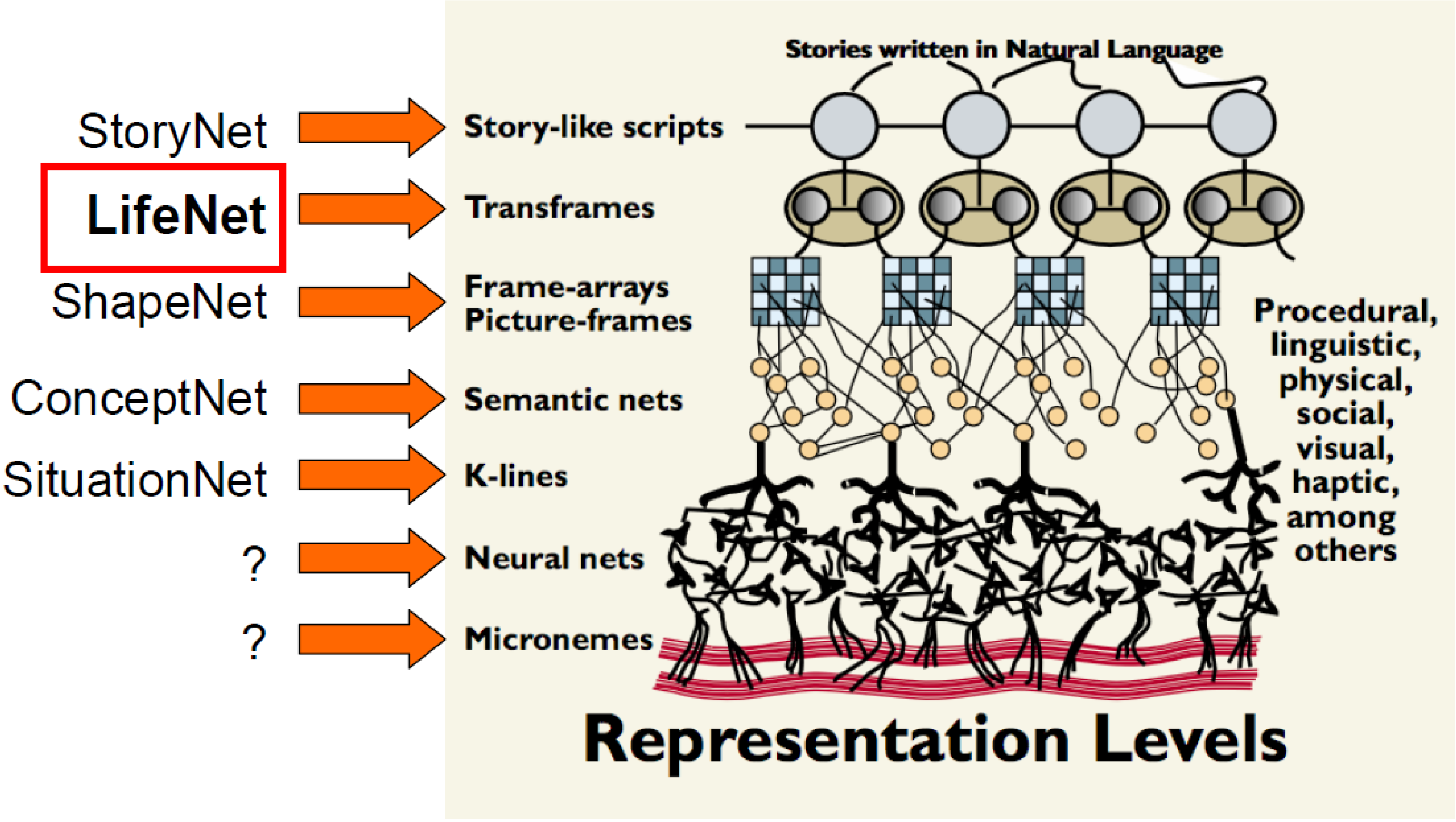


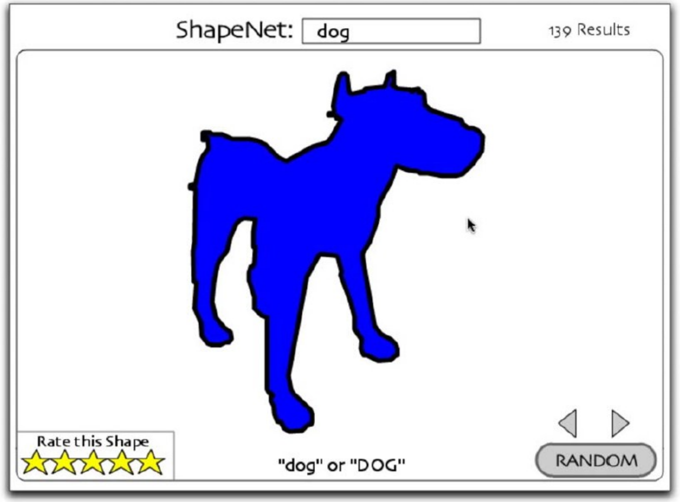
Fig 3 A sample of LifeNet. The before column shows t1 and the after column shows t2. 'It is 8 am' occurs before 'It is 11 am'. 'It is 8 am' occurs at the same time as 'I am brushing my teeth'.



Singh, Barbara Barry, and Hugo Liu (2004). **Teaching machines about everyday life.** *BT Technology Journal*, 22(4):227-240.

Representing Knowledge in Multiple Ways

ShapeNet: Spatial Common Sense



StoryNet

LifeNet

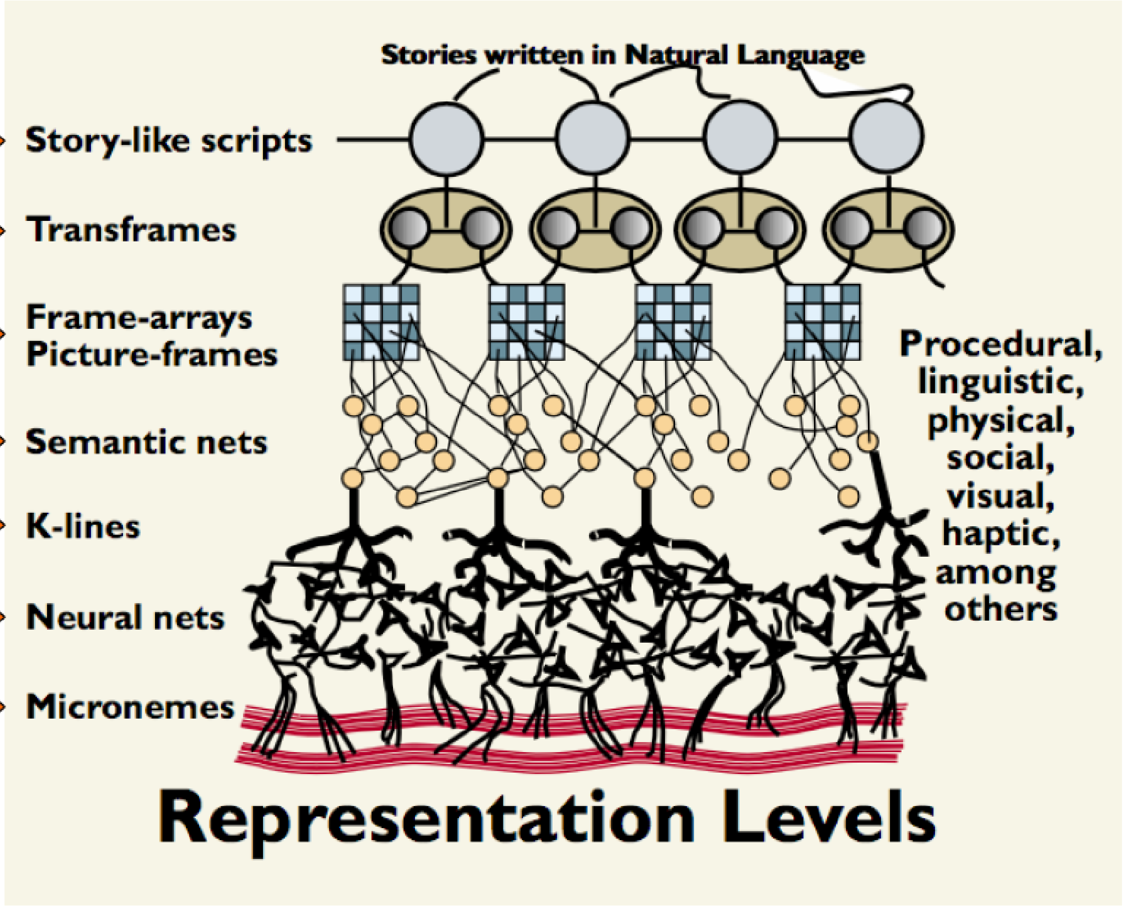
ShapeNet

ConceptNet

SituationNet

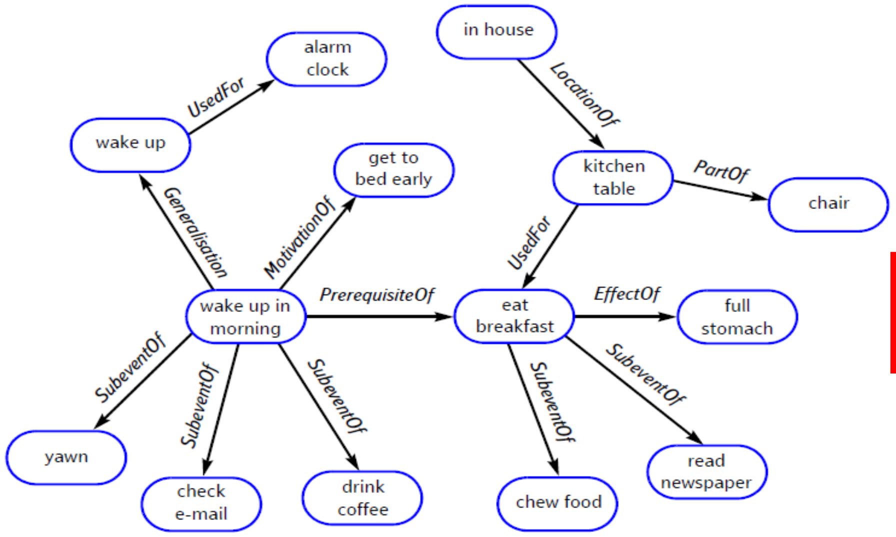
?

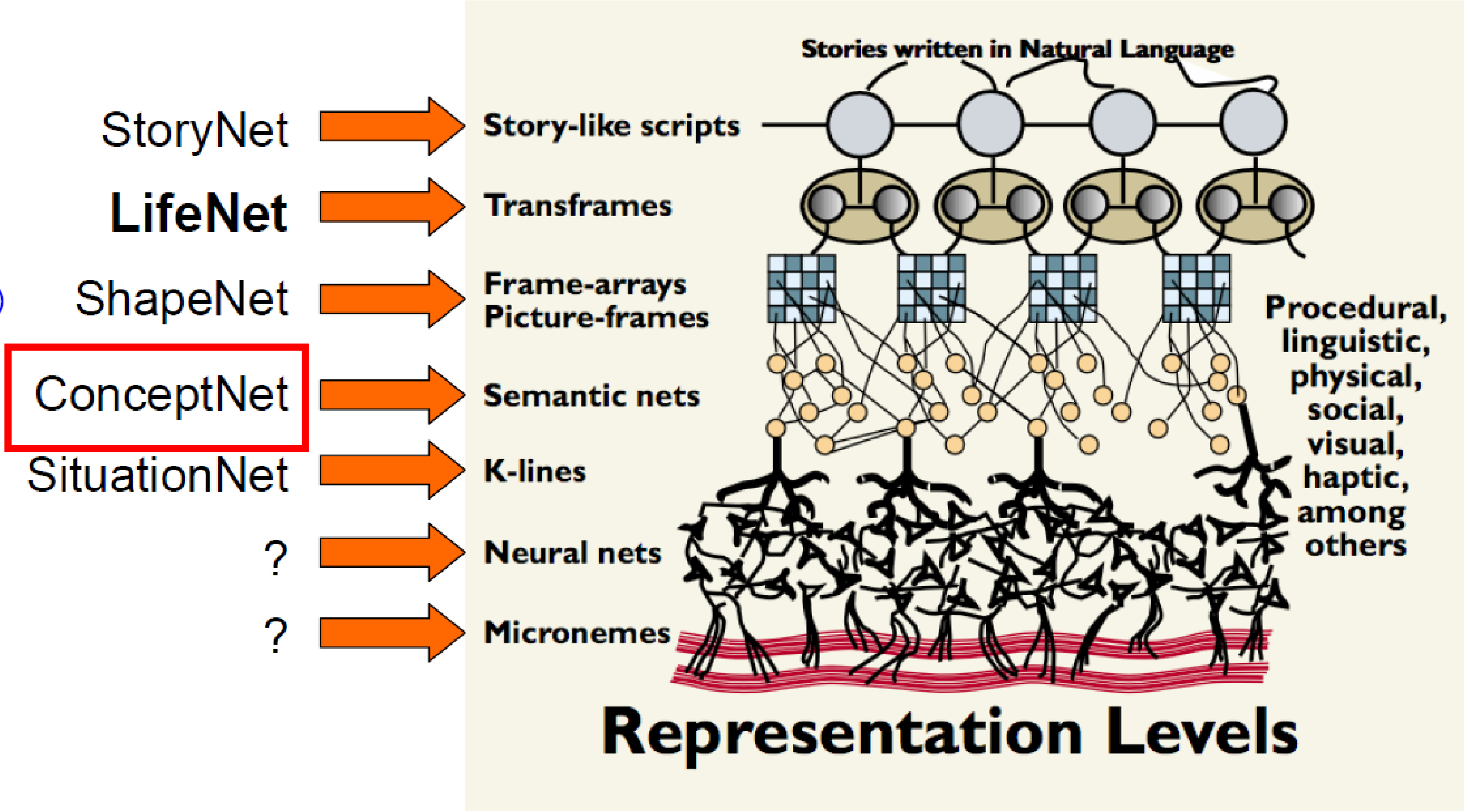
?



Singh, Barbara Barry, and Hugo Liu (2004). **Teaching machines about everyday life.** *BT Technology Journal*, 22(4):227-240.

Representing Knowledge in Multiple Ways





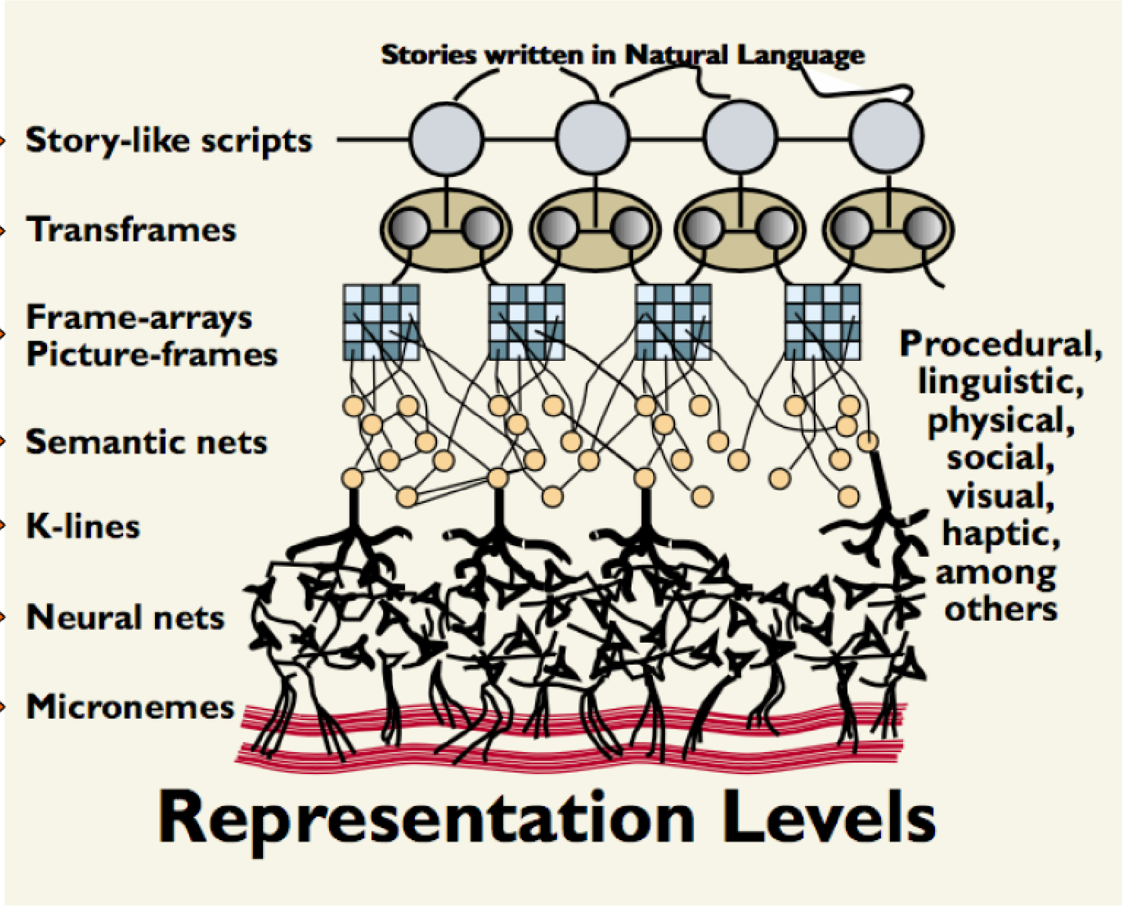
Singh, Barbara Barry, and Hugo Liu (2004). **Teaching machines about everyday life.** *BT Technology Journal*, 22(4):227-240.

Representing Knowledge in Multiple Ways

SituationNet:
detailed descriptions of situations

Prototypical situations

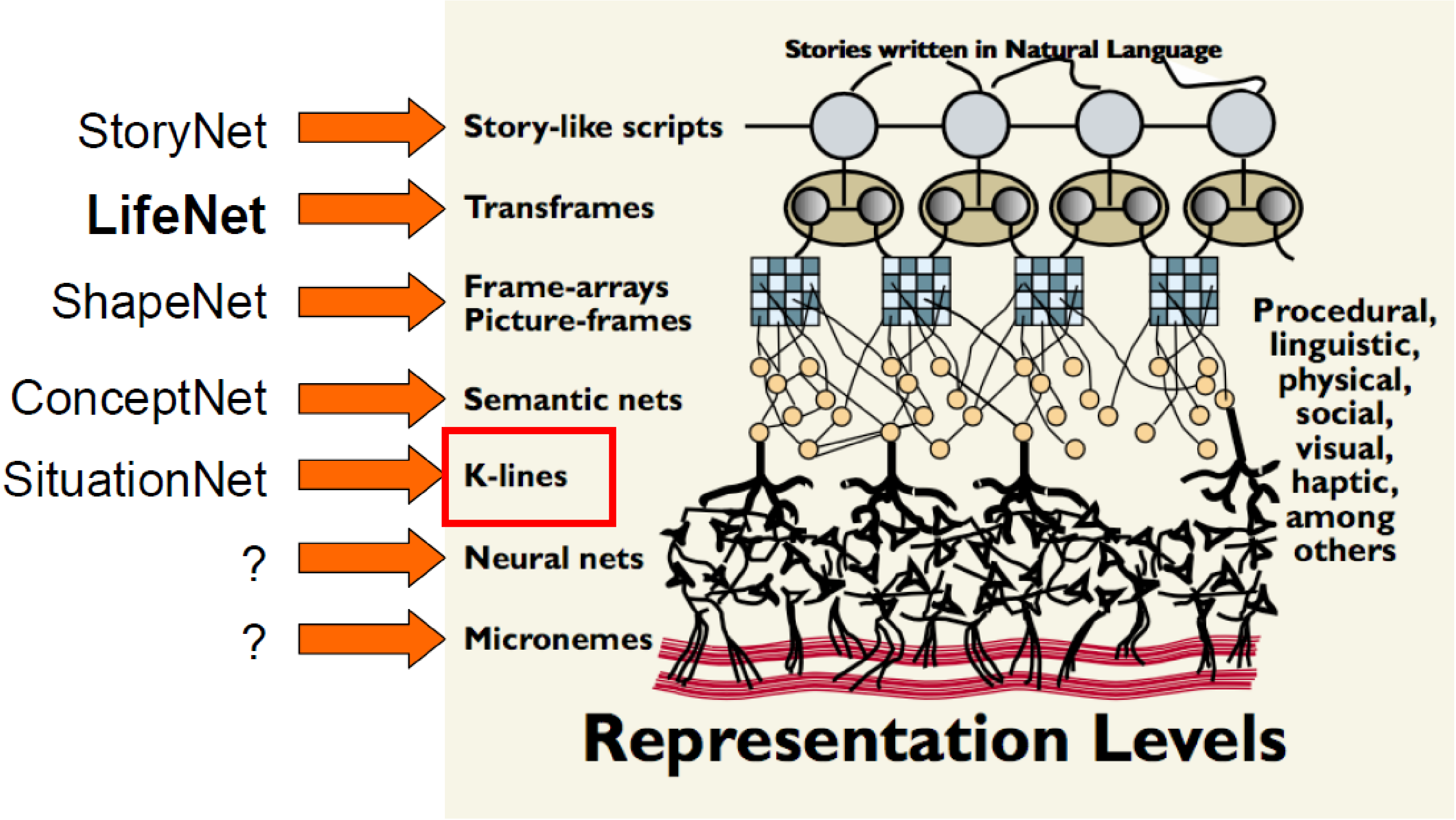
StoryNet
LifeNet
ShapeNet
ConceptNet
SituationNet
?
?



Singh, Barbara Barry, and Hugo Liu (2004). **Teaching machines about everyday life.** *BT Technology Journal*, 22(4):227-240.

Representing Knowledge in Multiple Ways

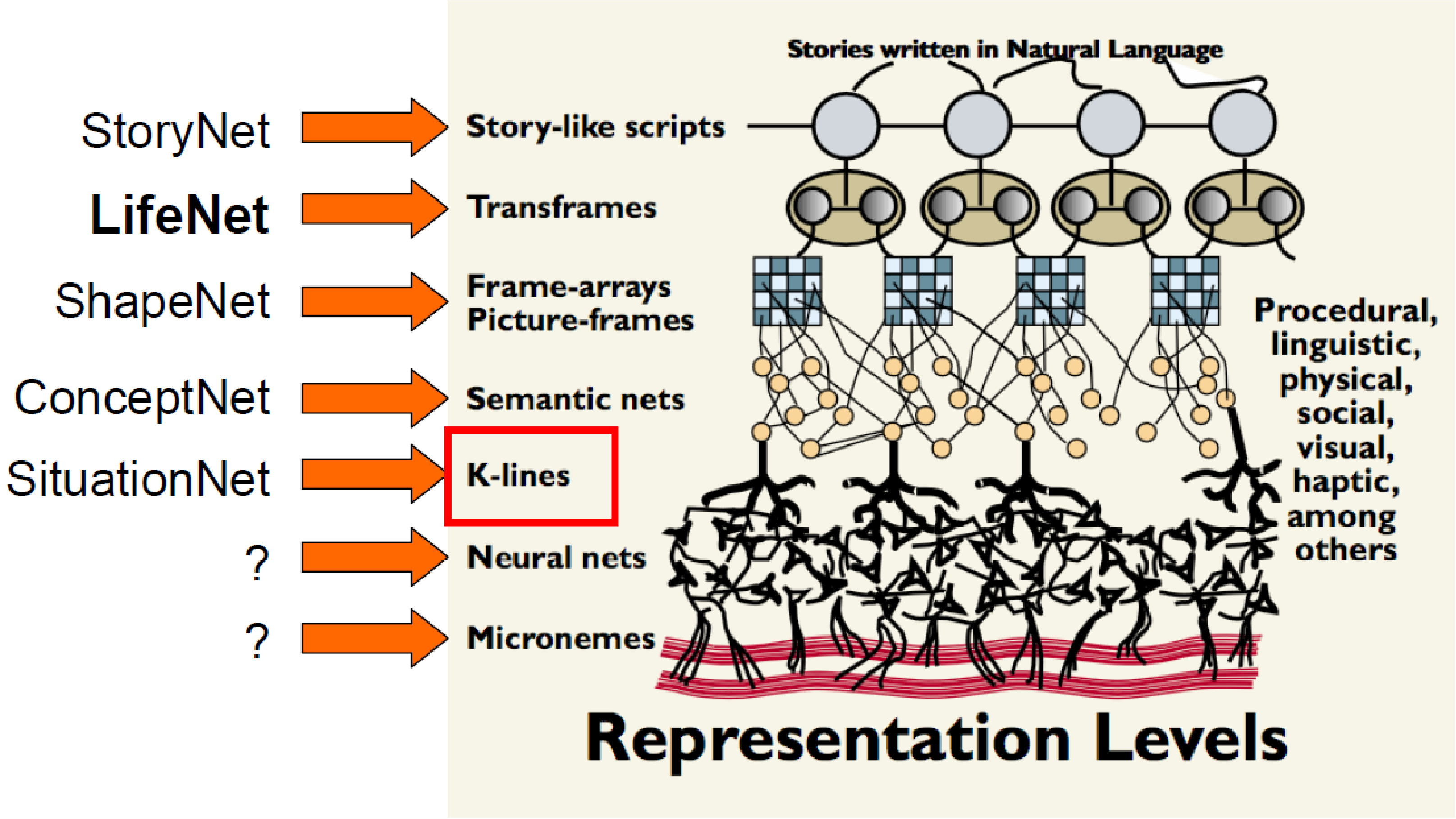
- "When you get an idea and want to "remember" it, you create a K-line for it."
- "When later activated, the K-line induces a partial mental state resembling the partial mental state that created that K-line."
- "A partial mental state is a subset of those mental agencies operating at one moment."



M. Minsky, "K-Lines: A theory of Memory," Cognitive Science 4 (1980). 117-133.

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".



M. Minsky, "K-Lines: A theory of Memory," Cognitive Science 4 (1980). 117-133.

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

Induction

X: Item does not fit in container, REASON, item is big

Deduction

Trophy is an item; Suitcase is a container

Conceptualization

Instantiation

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

Grounding

Consistent

CSKB/Training Data

- Computer not fit in parcel, REASON, Computer is big
- Rock not fit in carrier, REASON, rock is big
- …

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.

If we instantiate all, it's possible to entail

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)

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

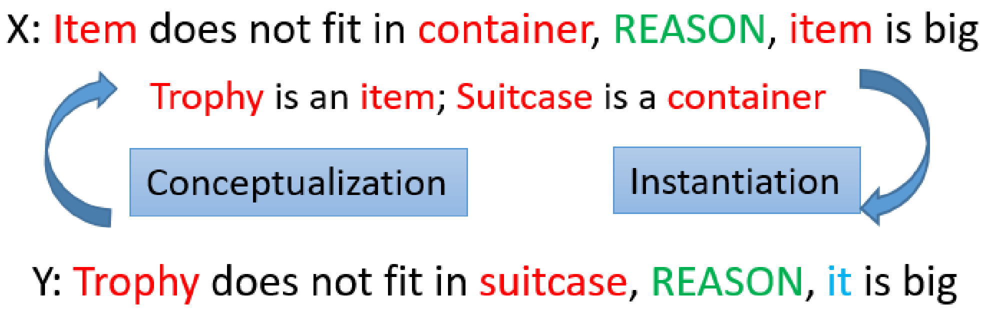
to get some beverage 

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

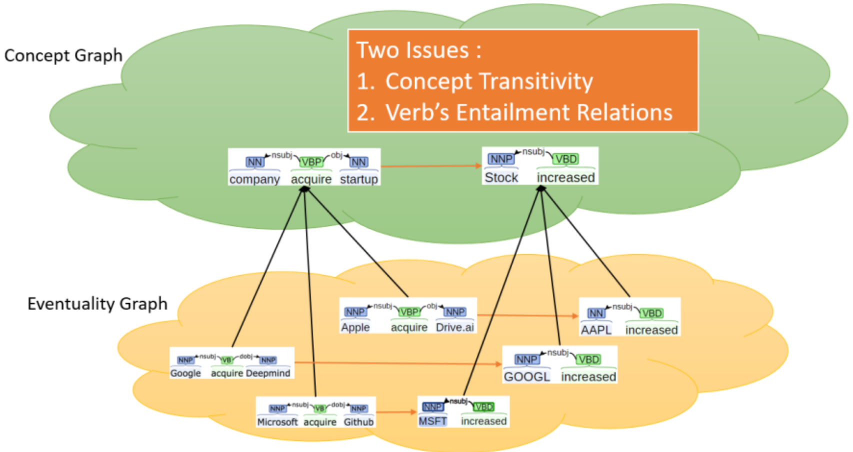


M. Minsky, “K-Lines: A theory of Memory," Cognitive Science 4 (1980). 117-133.

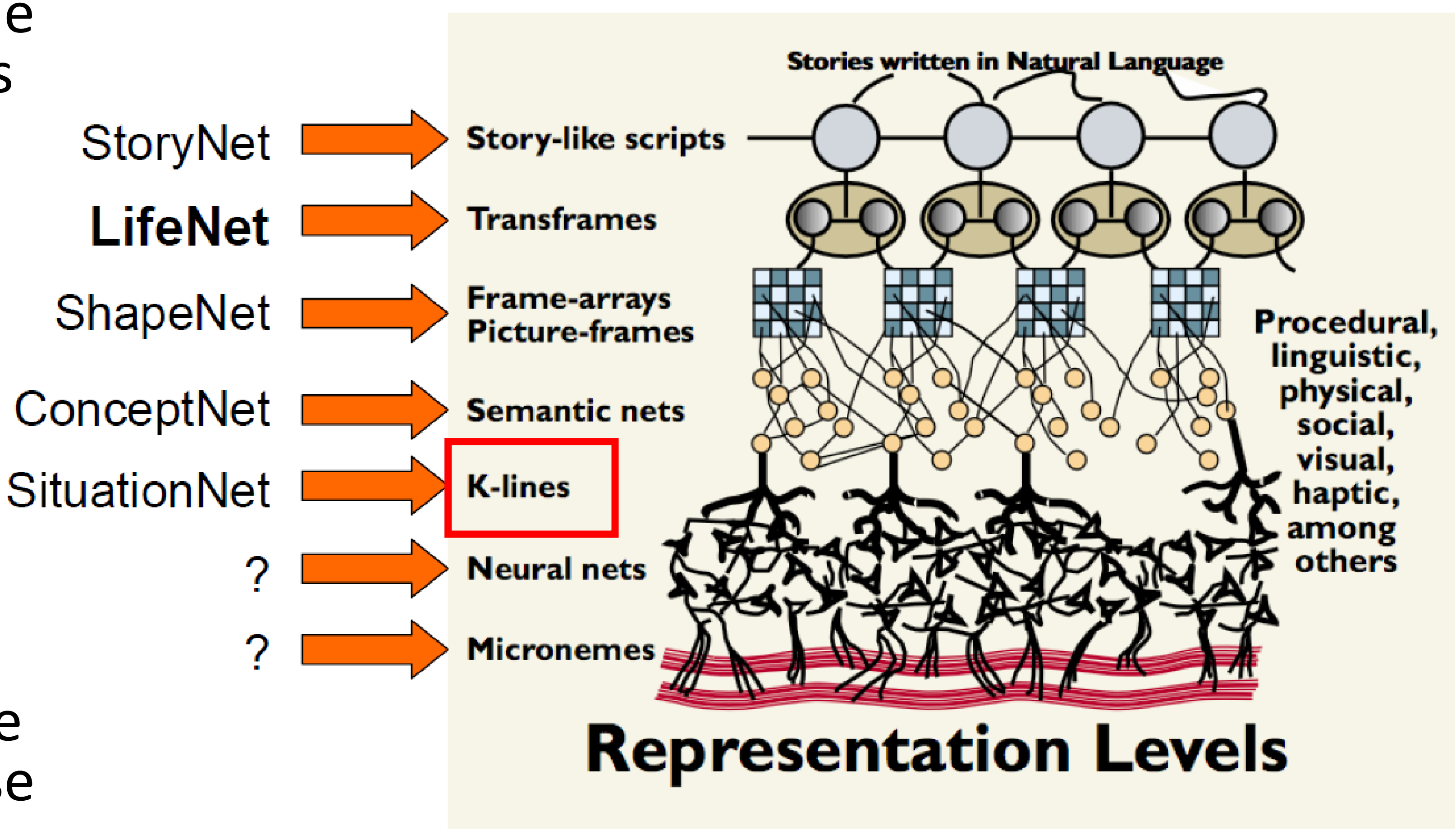
Representing Knowledge in Multiple Ways

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

ASER 2.0



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



M. Minsky, “K-Lines: A theory of Memory," *Cognitive Science 4 (1980). 117-133.*

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?

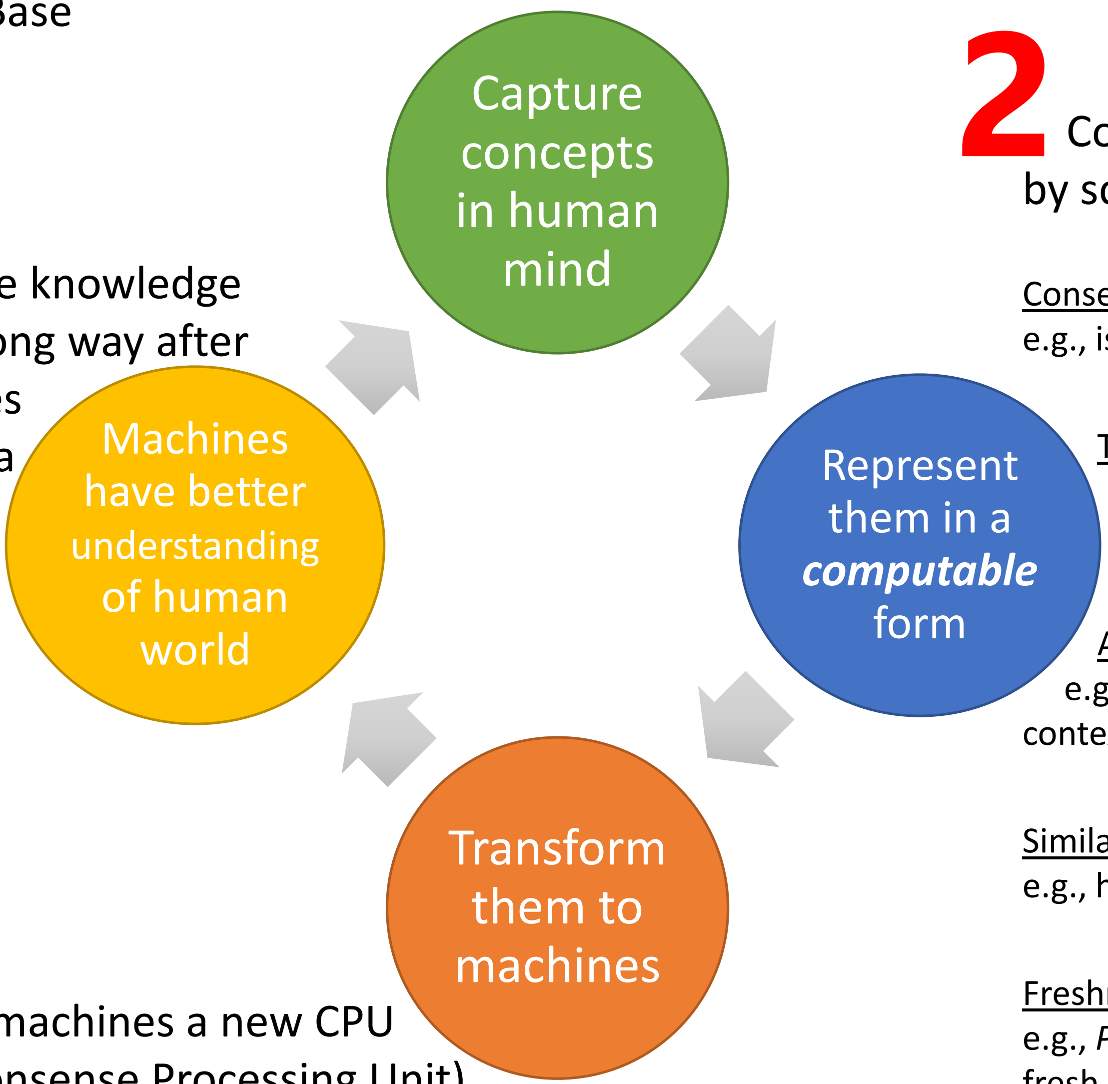
Typicality:
e.g. how likely you think of Apple when you think about companies?

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*.

…



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

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

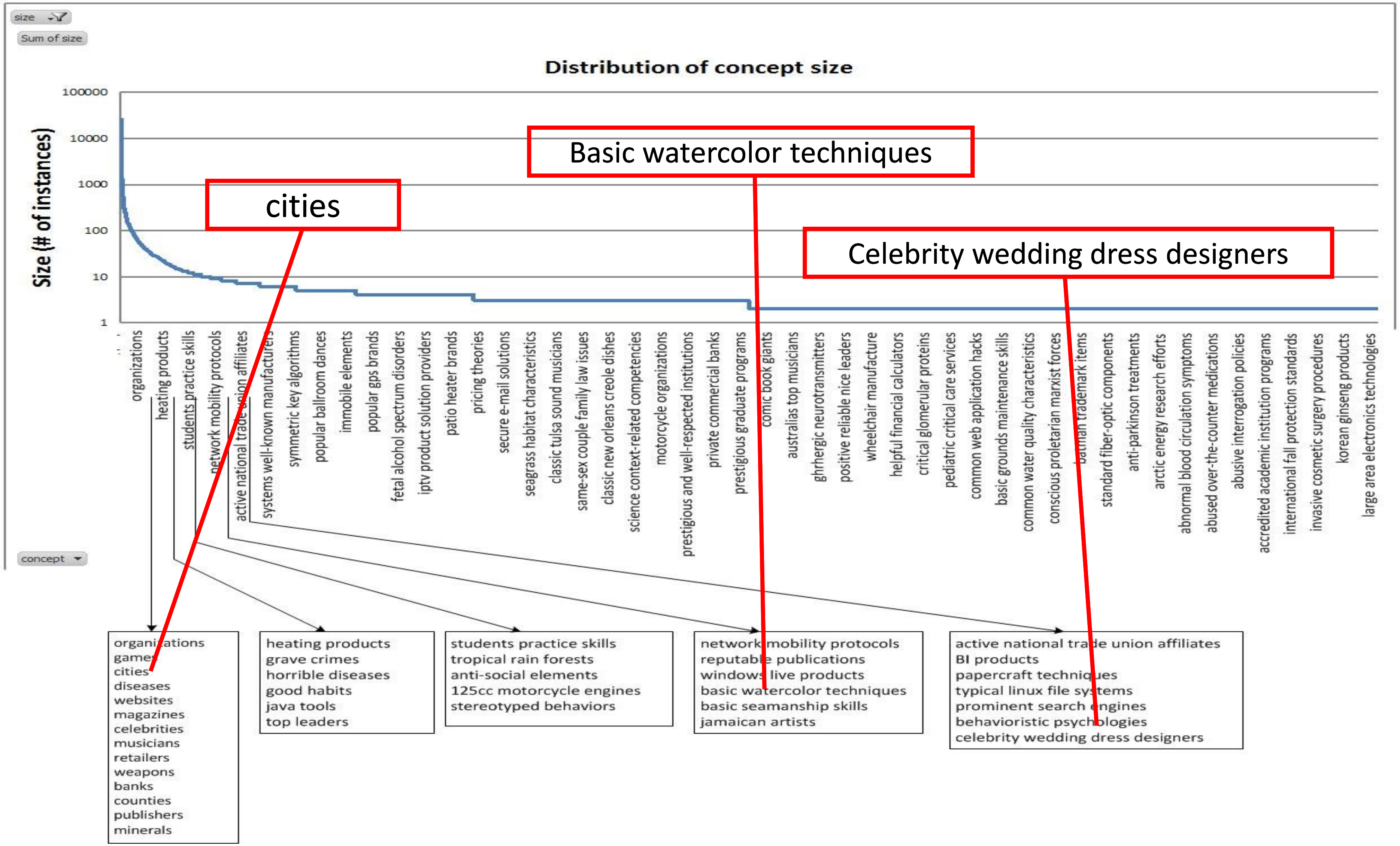
Slide Credit: Haixun Wang

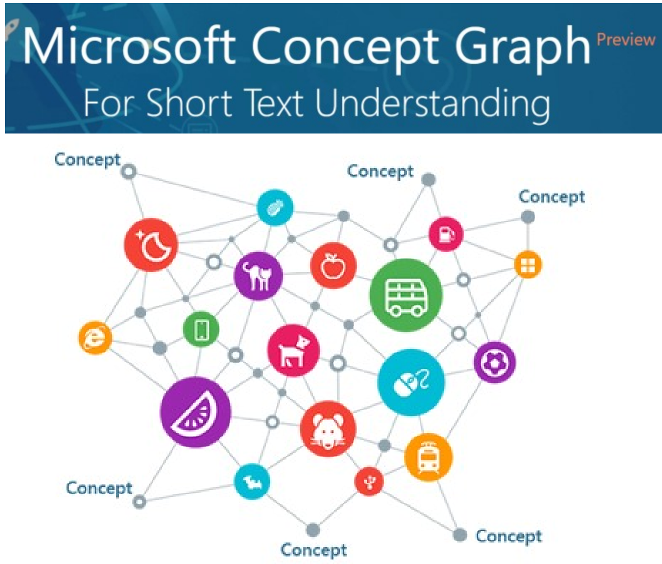
Data Sources

- 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)

Changlong Yu, Jialong Han, Peifeng Wang, Yangqiu Song, Hongming Zhang, Wilfred Ng, and Shuming Shi. When Hearst Is not Enough: Improving Hypernymy Detection from Corpus with Distributional Models. EMNLP. 2020.

ProBase





Probase is a *large, universal, probabilistic* knowledge base with **an extremely large concept space**

Data are available at https://concept.research.microsoft.com/

Wentao Wu, Hongsong Li, Haixun Wang, Kenny Qili Zhu: Probase: a probabilistic taxonomy for text understanding. SIGMOD Conference 2012: 481-492

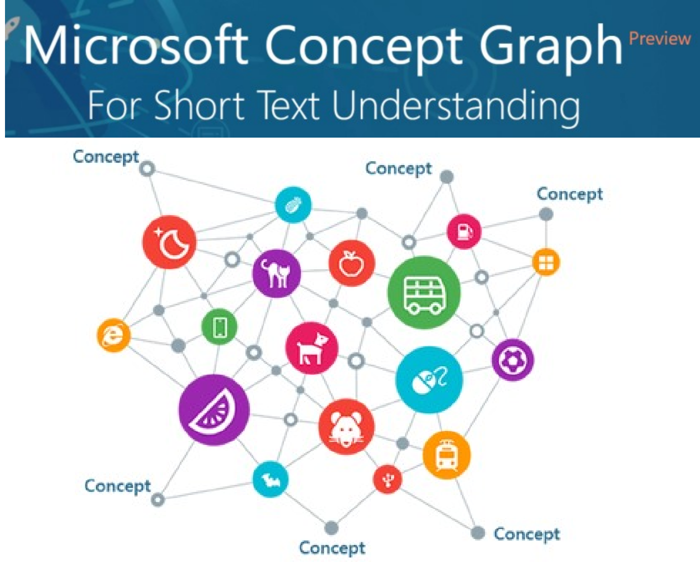
Slide Credit: Haixun Wang

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

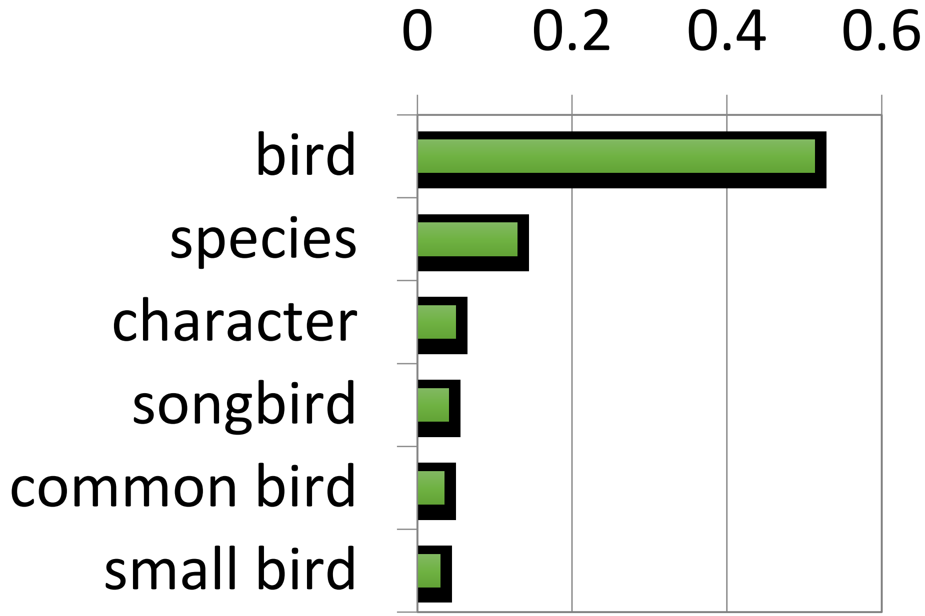
Slide Credit: Haixun Wang

Conceptualization with ProBase

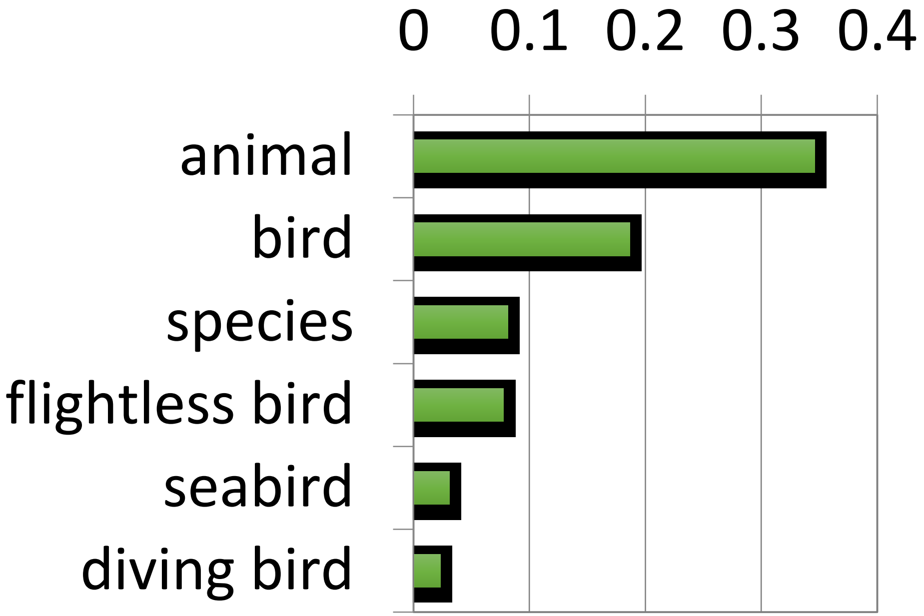


Typicality $$P(concept \mid instance) = \frac{\#(concept, instance)}{\#(instance)}$$

- Robin



- Penguin



Data are available at https://concept.research.microsoft.com/

Wentao Wu, Hongsong Li, Haixun Wang, Kenny Qili Zhu: Probase: a probabilistic taxonomy for text understanding. SIGMOD Conference 2012: 481-492
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.

How about others rather than entities and relations?

Jackendoff, R. (Ed.). (1990). Semantic Structures. Cambridge, Massachusetts: MIT Press.

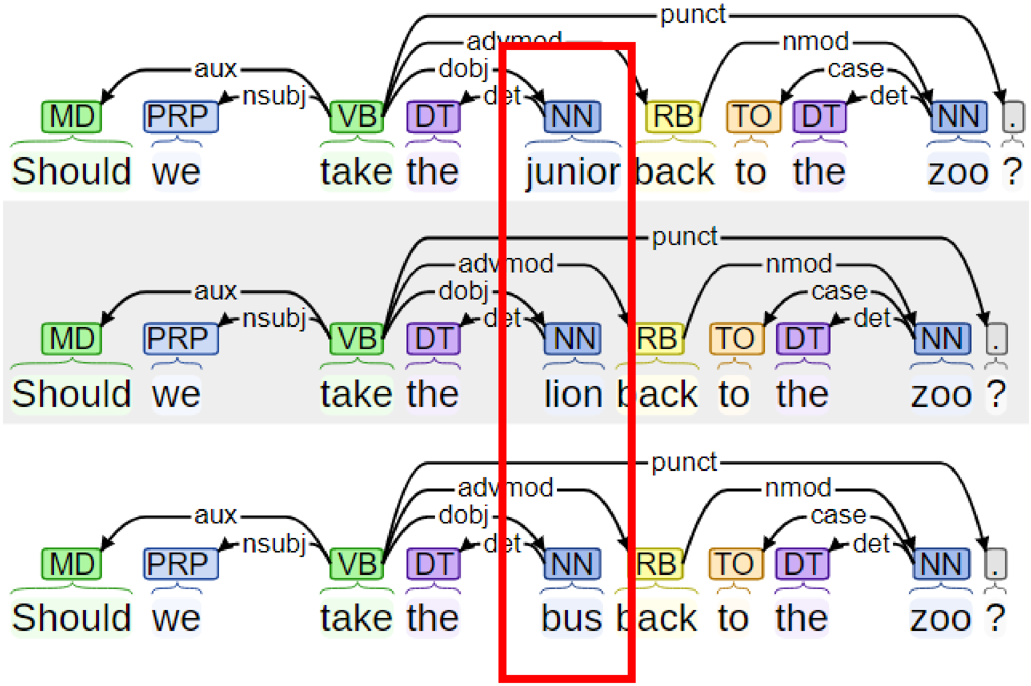
Semantic Primitive Units

- 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



Principle #1

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. Language, 39(2), 170–210.

Selectional Preference (SP)

Principle #2

- 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

Yorick Wilks. 1975. An intelligent analyzer and understander of English. Communications of the ACM, 18(5):264–274.
Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. Language, 39(2), 170–210.
Noam Chomsky. 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.
Philip Resnik. 1993. Selection and information: A class-based approach to lexical relationships. Ph.D. thesis, University of Pennsylvania.

A Test of Commonsense Reasoning

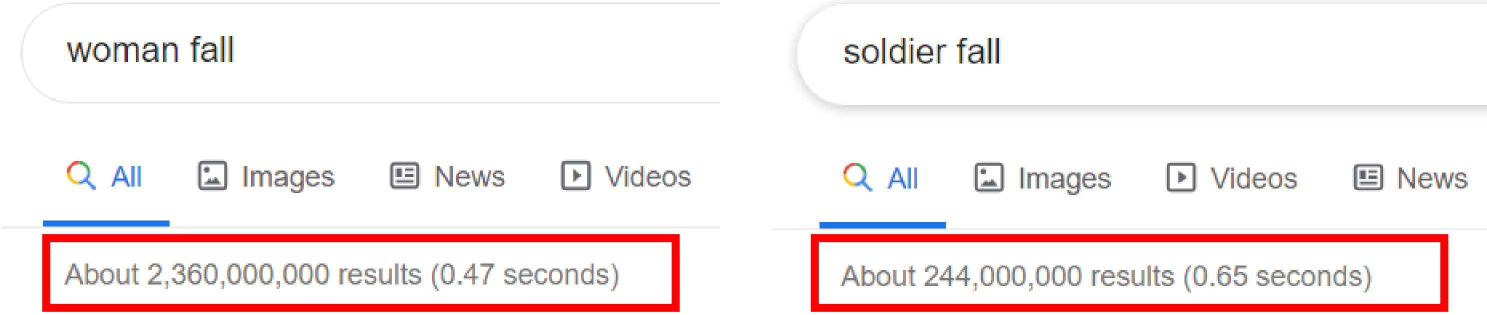
- Proposed by Hector Levesque at U of Toronto
- An example taking from Winograd Schema Challenge

  - (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

http://commonsensereasoning.org/winograd.html
https://en.wikipedia.org/wiki/Winograd_Schema_Challenge

The soldiers fired at the women, and we saw several of them fall.

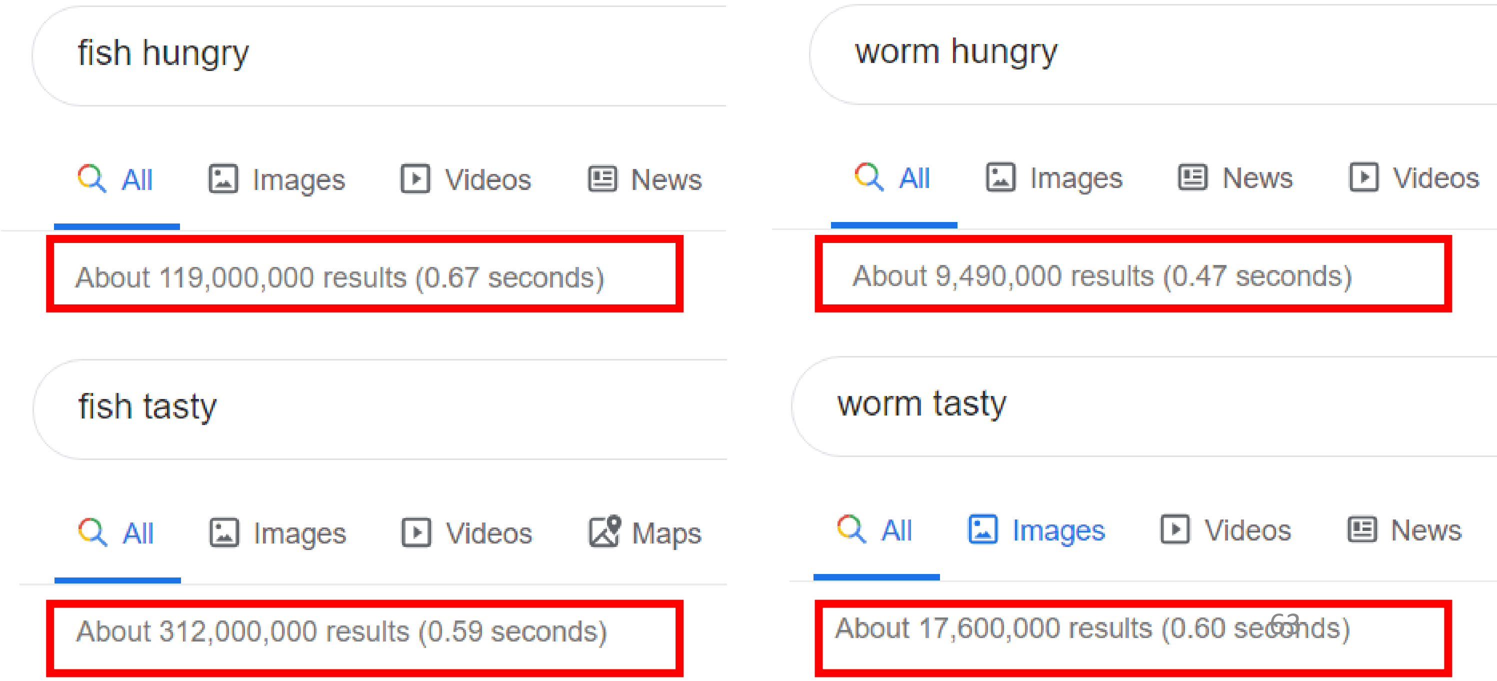


Why is it a challenge?

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



A Brief History of Datasets and Development

- Human's performance: 92.1% (Bender 2015)
- WinoGrande (RoBERTa + 43K Training data): 90.1% (Sakaguchi et al., 2019)

2011: Levesque. AAAI Spring Symposium

2012: The first large dataset. Rahman and Ng: EMNLP-CoNLL

2014: Davis et al. "A Collection of Winograd Schemas"

Recent results (Unsupervised/few-shot)

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

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

Hongming Zhang, Hantian Ding, and Yangqiu Song. SP-10K: A Large-Scale Evaluation Set for Selectional Preference Acquisition. ACL, 2019.

Examples in SP-10K

dobj	Plausibility
(eat, meal)	10.00
(close, door)	8.50
(touch, food)	5.50
(hate, investment)	4.00
(eat, mail)	0.00

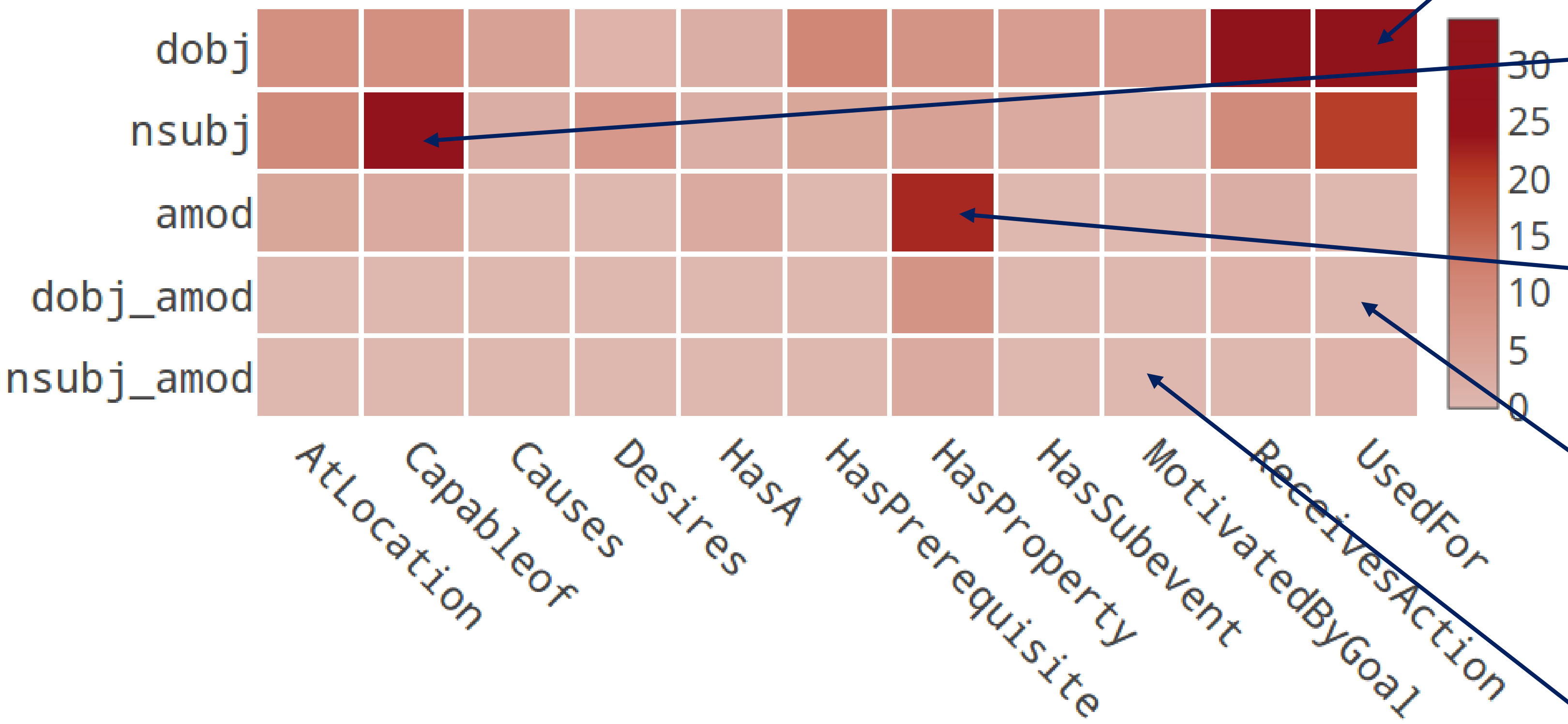
nsubj	Plausibility
(singer, sing)	10.00
(law, permit)	7.78
(women, pray)	5.83
(victim, contain)	2.22
(textbook, eat)	0.00

amod	Plausibility
(fresh, air)	9.77
(new, method)	8.89
(medium, number)	4.09
(immediate, food)	2.05
(secret, wind)	0.75

dobj_amod	Plausibility
(lift, heavy *object*)	9.17
(design, new *object*)	8.00
(attack, small *object*)	5.23
(inform, weird *object*)	3.64
(earn, rubber *object*)	0.63

nsubj_amod	Plausibility
(evil *subject*, attack)	9.00
(recent *subject*, demonstrate)	6.00
(random *subject*, bear)	4.00
(happy *subject*, steal)	2.25
(sunny *subject*, make)	0.56

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

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

dobj_amod	Plausibility
(lift, heavy *object*)	9.17
(design, new *object*)	8.00
(attack, small *object*)	5.23
(inform, weird *object*)	3.64
(earn, rubber *object*)	0.63

nsubj_amod	Plausibility
(evil *subject*, attack)	9.00
(recent *subject*, demonstrate)	6.00
(random *subject*, bear)	4.00
(happy *subject*, steal)	2.25
(sunny *subject*, make)	0.56

*PP: posterior probability for SP acquisition using Wikipedia data

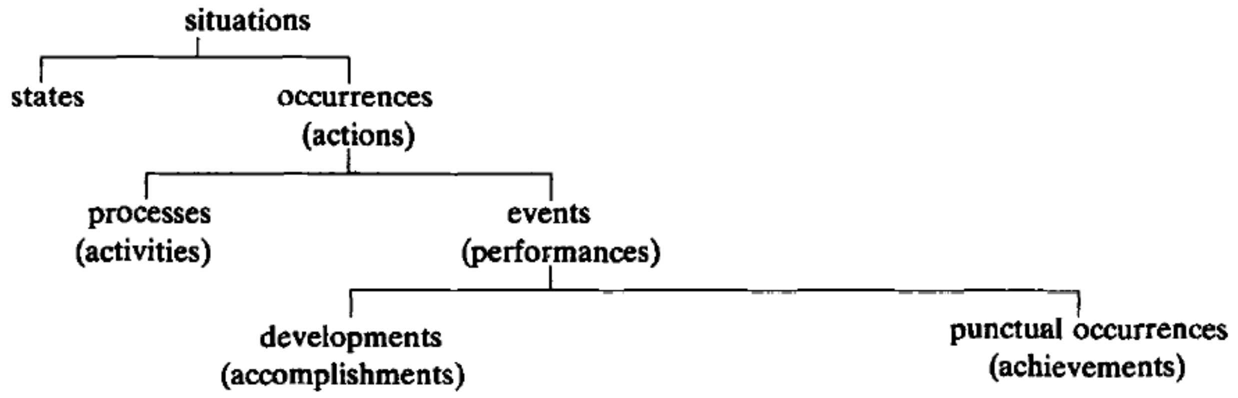
KnowlyWood

- Perform information extraction from free text
  - Mostly movie scripts and novel books
- Four relations: previous, next, parent, similarity
- No subject information
  - Only verb+object

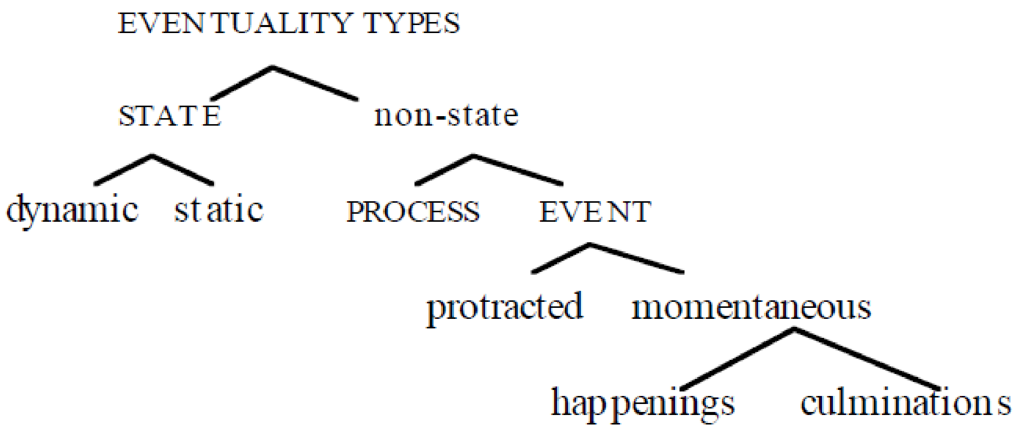
Niket Tandon, Gerard de Melo, Abir De, Gerhard Weikum: Knowlywood: Mining Activity Knowledge From Hollywood Narratives. CIKM 2015: 223-232

ASER (**A**ctivities, **S**tates, **E**vents, and their **R**elations)

Mourelatos' taxonomy (1978)



Bach's taxonomy (1986)



- **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.

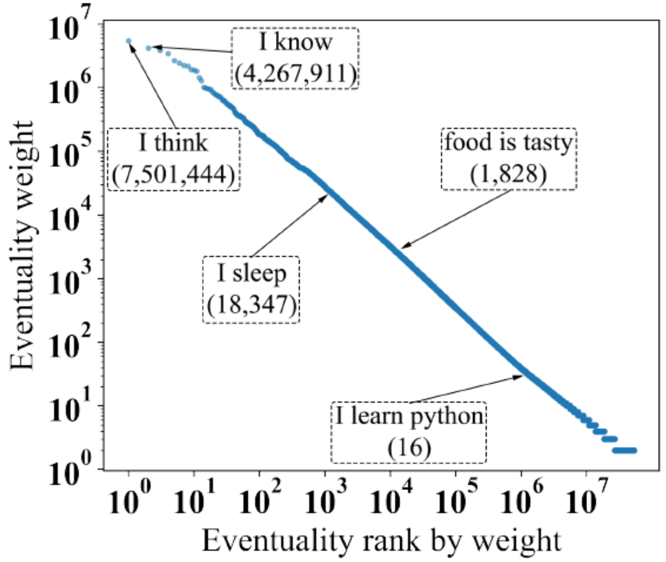
- **Static states**: be in New York, love (one's cat);
- **Dynamic states**: sit, stand, drunk, present, sick;
- **Processes**: walk, push a cart, sleep;
- **Protracted events**: build (a cabin), eat a sandwich, polish a shoe, walk to Boston;
- **Culminations**: take off; arrive, leave, depart;
- **Happenings**: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

Alexander P. D. Mourelatos. Events, processes, and states. Linguistics and Philosophy, 2, 415-434. 1978.
Emmon Bach. The algebra of events. Linguistics and philosophy, 9 (1), 5-16. 1986.

Eventualities

- Using patterns to collect partial information
- Six relations are also kept but treated as auxiliary edges
  - advmod,
  - amod,
  - nummod,
  - aux,
  - compound,
  - neg

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



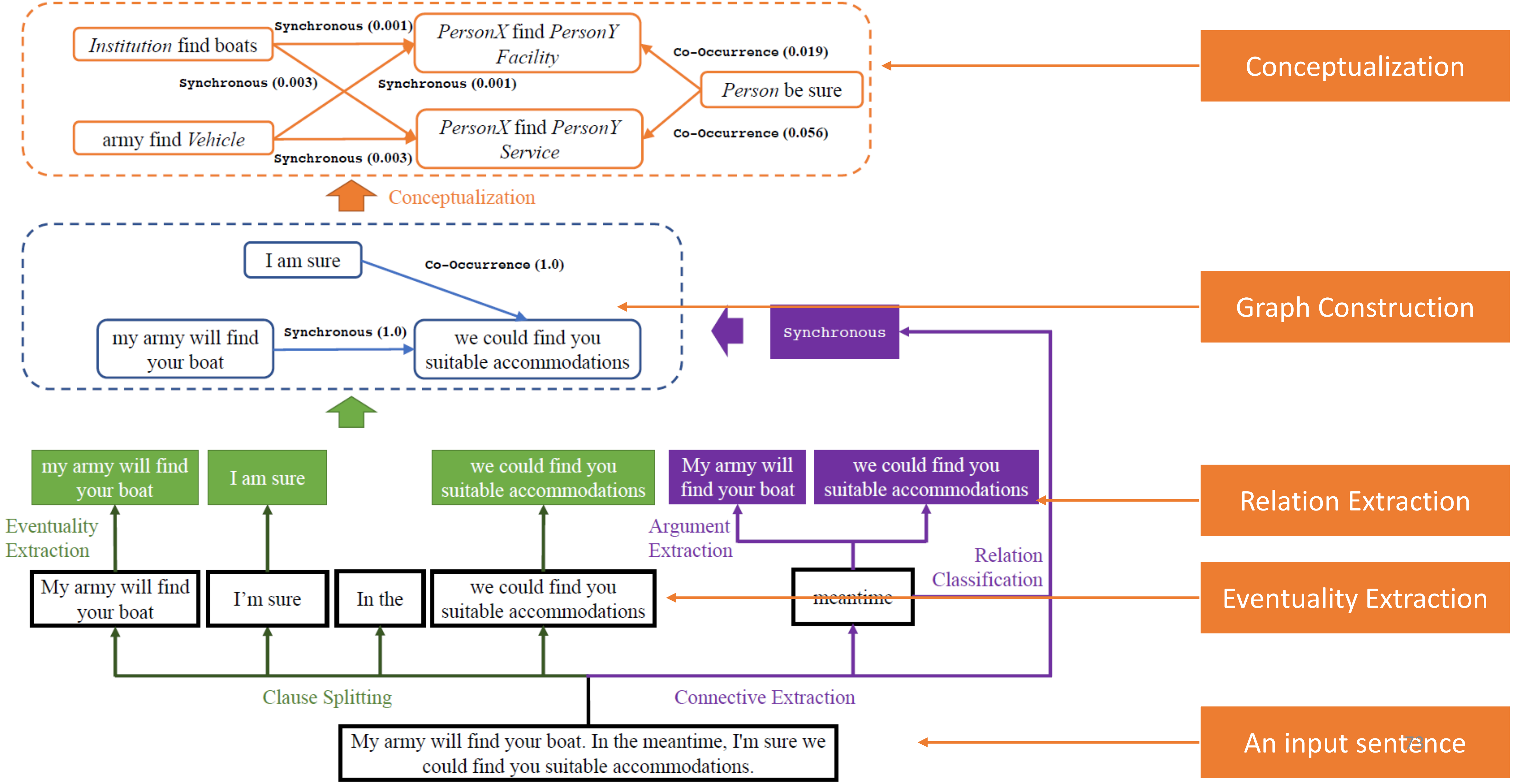
Eventuality Relations

- 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)

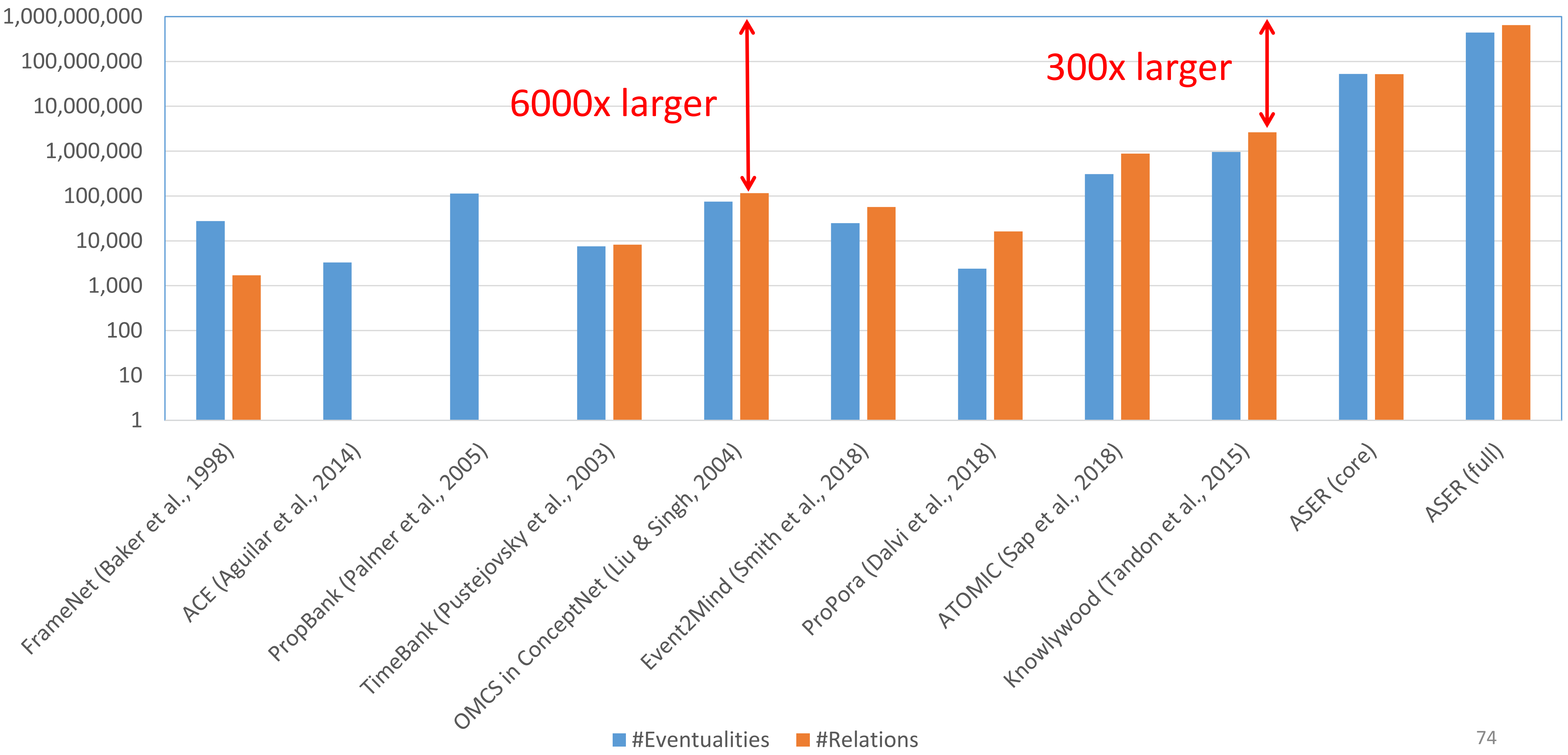
Relation Type	Examples
Precedence	E1 **before** E2; E1 , **then** E2; E1 **till** E2; E1 **until** E2
Succession	E1 **after** E2; E1 **once** E2
Synchronous	E1, **meanwhile** E2; E1 **meantime** E2; E1, **at the same time** E2
Reason	E1, **because** E2
Result	E1, **so** E2; E1, **thus** E2; E1, **therefore** E2; E1, **so that** E2
Condition	E1, **if** E2; E1, **as long as** E2
Contrast	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
Concession	E1, **although** E2
Conjunction	E1 **and** E2; E1, **also** E2
Instantiation	E1, **for example** E2; E1, **for instance** E2
Restatement	E1, **in other words** E2
Alternative	E1 **or** E2; E1, **unless** E2; E1, **as an alternative** E2; E1, **otherwise** E2
ChosenAlternative	E1, E2 **instead**
Exception	E1, **except** E2

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., & Webber, B. L. (2007). The penn discourse treebank 2.0 annotation manual.
Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, Attapol T. Rutherford. The CoNLL-2015 Shared Task on Shallow Discourse Parsing.
Jianxiang Wang and Man Lan. A Refined End-to-End Discourse Parser. CONLL Shared Task 2015.

A Running Example



Scales of Verb Related Knowledge Graphs



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 concepualize?

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)

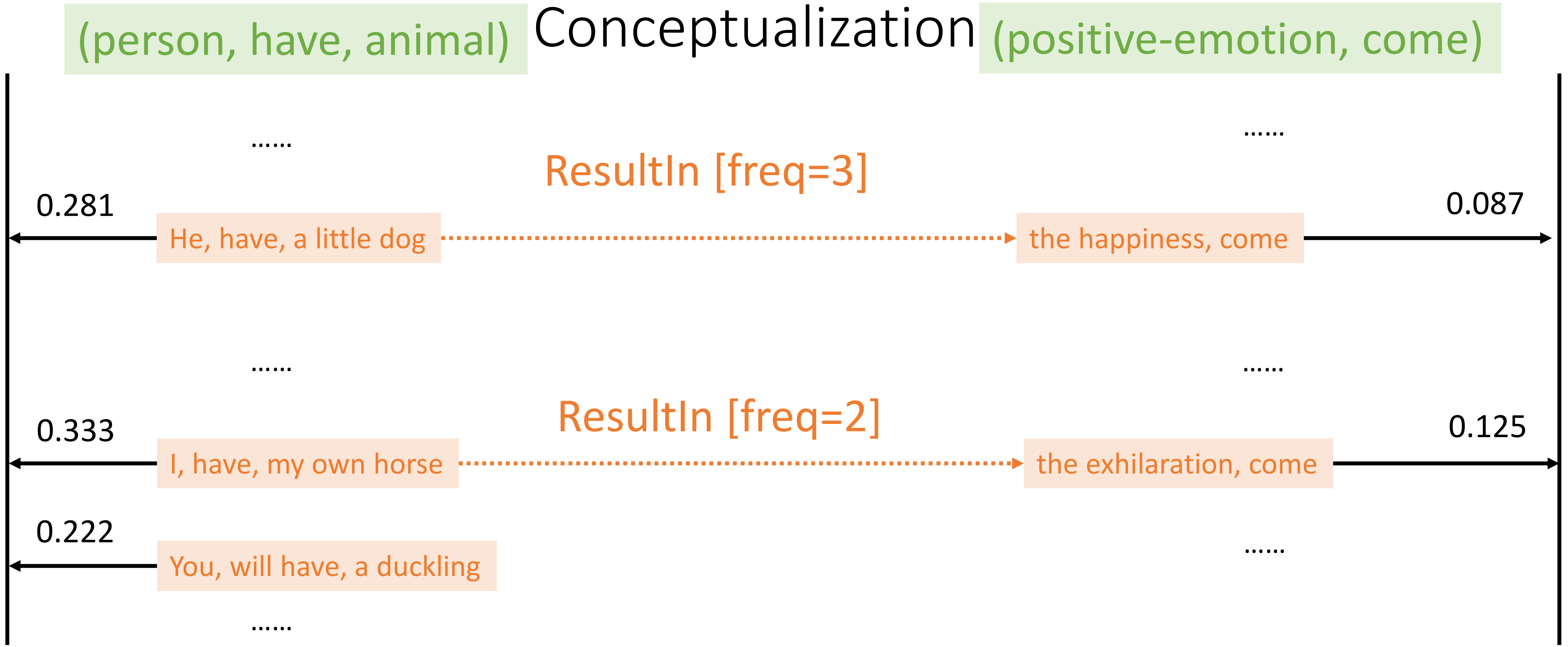
- “stocks price may increase when a company acquires a start-up”

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

Normalization

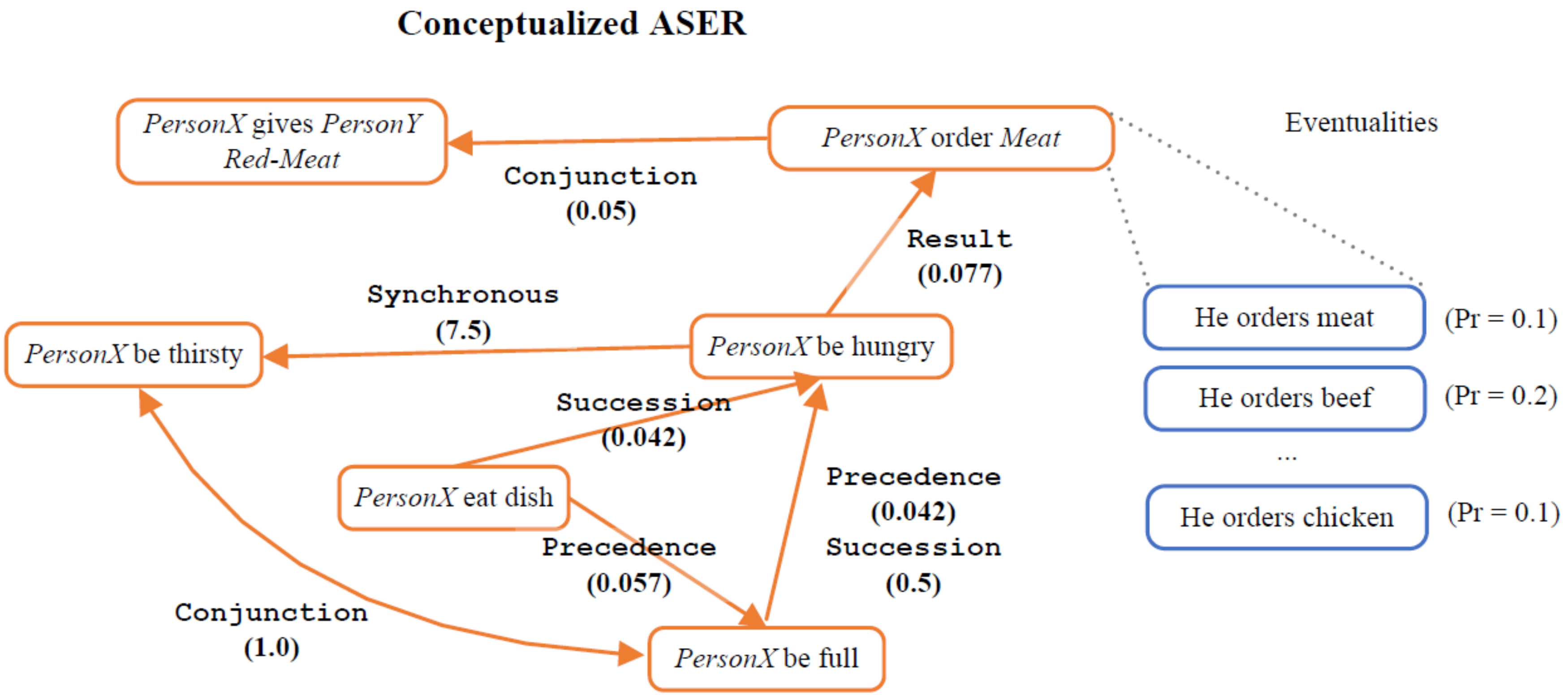
			Probability
He, she, I, Bob, …	$\longrightarrow$	PERSON	1.0
1996, 2020, 1949, …	$\longrightarrow$	YEAR	1.0
23, 20, 333, ….	$\longrightarrow$	DIGIT	1.0
www.google.com, …	$\longrightarrow$	URL	1.0

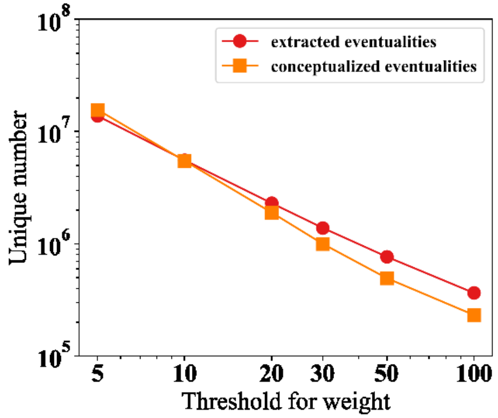
Conceptualization

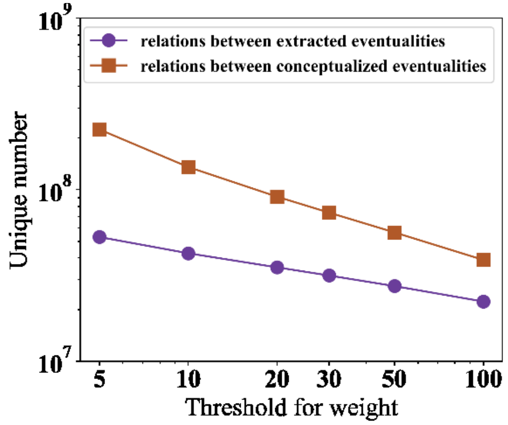


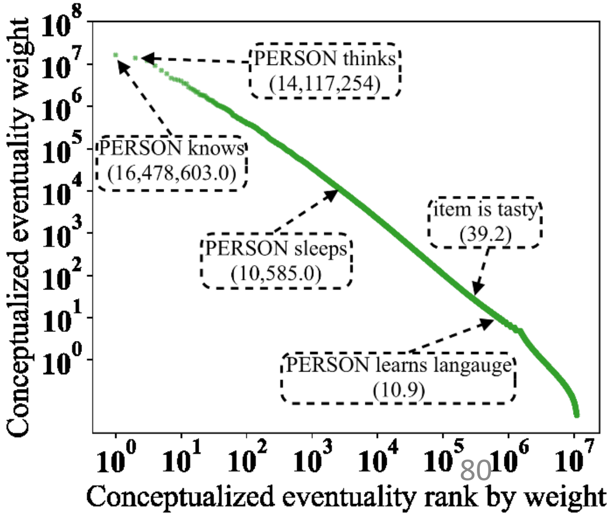
P(ResultIn | (person, have, animal) , (positive-emotion, come)) $= 0.281 \times 3 \times 0.087 + 0.333 \times 2 \times 0.125$
$= 0.157$

Conceptualization Examples









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)

Data	#Unique Eventualities	#Unique Relations
Core	34 millions	15 millions
Full	272 millions	206 millions

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

Data	#Unique Eventualities	#Unique Relations
Core	53 millions	52 millions
Full	439 millions	649 millions

- 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)

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)

Rule Mining: Eventualities

- Mine Rules using AMIE + $< E_a, T_1, E_b > \wedge < E_b, T_2, E_c > \Rightarrow < E_a, T_3, E_b >,$

Rule	$\langle E_b \xrightarrow{\texttt{Concession}} E_f \rangle \wedge \langle E_a \xrightarrow{\texttt{Result}} E_f \rangle \Rightarrow \langle E_a \xrightarrow{\texttt{Contrast}} E_b \rangle$	Concession: E1, **although** E2
Instances	⟨ I do not know → I guess ⟩ ∧ ⟨ I believe → I guess ⟩ ⇒ ⟨ I believe → I do not know ⟩ ⟨ I am not sure → I guess ⟩ ∧ ⟨ I hope so → I guess ⟩ ⇒ ⟨ I hope so → I am not sure ⟩ ⟨ I understand → I can not speak ⟩ ∧ ⟨ I am not a lawyer → I can not speak ⟩ ⇒ ⟨ I am not a lawyer → I understand ⟩	
Rule	$\langle E_f \xrightarrow{\texttt{Contrast}} E_b \rangle \wedge \langle E_a \xrightarrow{\texttt{Instantiation}} E_f \rangle \Rightarrow \langle E_a \xrightarrow{\texttt{Contrast}} E_b \rangle$	
Instances	⟨ I remember → I could not find it ⟩ ∧ ⟨ I get → I remember ⟩ ⇒ ⟨ I get → I could not find it ⟩ ⟨ I would say → I might be wrong ⟩ ∧ ⟨ I hope → I would say ⟩ ⇒ ⟨ I hope → I might be wrong ⟩ ⟨ It have been suggested → This is unlikely ⟩ ∧ ⟨ It is possible → It have been suggested ⟩ ⇒ ⟨ It is possible → This is unlikely ⟩	
Rule	$\langle E_e \xrightarrow{\texttt{ChosenAlternative}} E_b \rangle \wedge \langle E_a \xrightarrow{\texttt{ChosenAlternative}} E_e \rangle \Rightarrow \langle E_a \xrightarrow{\texttt{ChosenAlternative}} E_b \rangle$	ChosenAlternative: E1, E2 **instead**
Instances	⟨ I will not go → You come here ⟩ ∧ ⟨ I want to see → I will not go ⟩ ⇒ ⟨ I want to see → You come here ⟩ ⟨ I want → It is ⟩ ∧ ⟨ I wish → I want ⟩ ⇒ ⟨ I wish → It is ⟩ ⟨ I want → I get ⟩ ∧ ⟨ I do not get that → I want ⟩ ⇒ ⟨ I do not get that → I get ⟩	

Luis Galárraga, Christina TeflioudiFabian Suchanek, Katja Hose Fast Rule Mining in Ontological Knowledge Bases with AMIE+. VLDB Journal 2015.

Rule Mining: Concepts

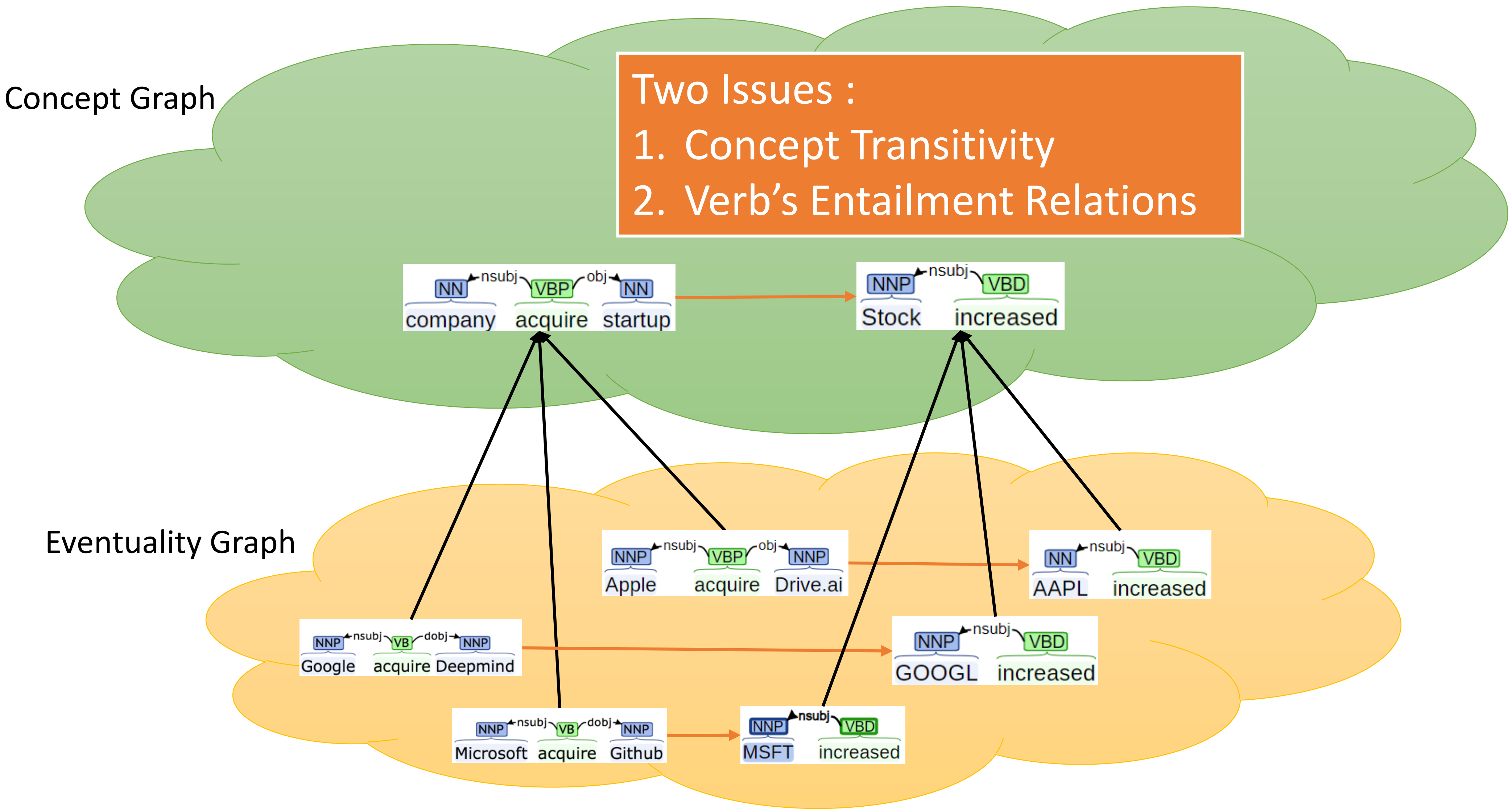
- Mine Rules using AMIE+ $< E_a, T_1, E_b > \wedge < E_b, T_2, E_c > \Rightarrow < E_a, T_3, E_b >,$

Rule	$\langle E_e \xrightarrow{\texttt{Restatement}} E_a \rangle \wedge \langle E_e \xrightarrow{\texttt{Restatement}} E_b \rangle \Rightarrow \langle E_a \xrightarrow{\texttt{Conjunction}} E_b \rangle$
Instances	⟨ *PersonX* laugh → *PersonX* smile ⟩ ∧ ⟨ *PersonX* laugh → *PersonX* open *Facial-Feature* ⟩ ⇒ ⟨ *PersonX* smile → *PersonX* open *Facial-Feature* ⟩
	⟨ *PersonX* love it → It be good ⟩ ∧ ⟨ *PersonX* love it → It be tasty ⟩ ⇒ ⟨ It be good → It be tasty ⟩
	⟨ *PersonX* wish → *PersonX* need ⟩ ∧ ⟨ *PersonX* wish → *PersonX* need ⟩ ⇒ ⟨ *PersonX* need → *PersonX* need ⟩
Rule	$\langle E_e \xrightarrow{\texttt{Instantiation}} E_a \rangle \wedge \langle E_e \xrightarrow{\texttt{Instantiation}} E_b \rangle \Rightarrow \langle E_a \xrightarrow{\texttt{Conjunction}} E_b \rangle$
Instances	⟨ *PersonX* realize → *PersonX* point out ⟩ ∧ ⟨ *PersonX* realize → PersonX have *Information* ⟩ ⇒ ⟨ *PersonX* point out → *PersonX* have *Information* ⟩
	⟨ *PersonX* have → *PersonX* get ⟩ ∧ ⟨ *PersonX* have → *PersonX* own ⟩ ⇒ ⟨ *PersonX* get → *PersonX* own ⟩
	⟨ *PersonX* know → *PersonX* be sure ⟩ ∧ ⟨ *PersonX* know → *PersonX* remember ⟩ ⇒ ⟨ *PersonX* be sure → *PersonX* remember ⟩
Rule	$\langle E_e \xrightarrow{\texttt{Concession}} E_b \rangle \wedge \langle E_e \xrightarrow{\texttt{Restatement}} E_a \rangle \Rightarrow \langle E_a \xrightarrow{\texttt{Contrast}} E_b \rangle$
Instances	⟨ *PersonX* order *Dish* → *PersonX* be hungry ⟩ ∧ ⟨ *PersonX* order *Dish* → *PersonX* order ⟩ ⇒ ⟨ *PersonX* order → *PersonX* be hungry ⟩
	⟨ *PersonX* wish → *PersonX* doubt ⟩ ∧ ⟨ *PersonX* wish → *PersonX* need ⟩ ⇒ ⟨ *PersonX* doubt → *PersonX* need ⟩
	⟨ *PersonX* love it → *PersonX* hate it ⟩ ∧ ⟨ *PersonX* love it → It be good ⟩ ⇒ ⟨ *PersonX* hate it → It be good ⟩

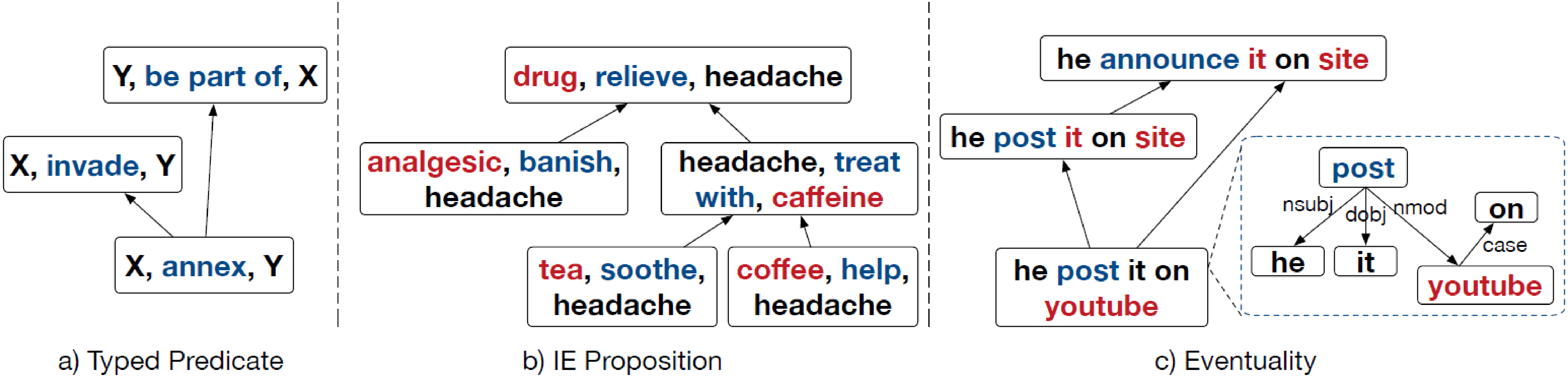
Instantiation	E1, **for example** E2; E1, **for instance** E2
Restatement	E1, **in other words** E2

Luis Galárraga, Christina TeflioudiFabian Suchanek, Katja Hose Fast Rule Mining in Ontological Knowledge Bases with AMIE+. VLDB Journal 2015.

Incorporating More Relations

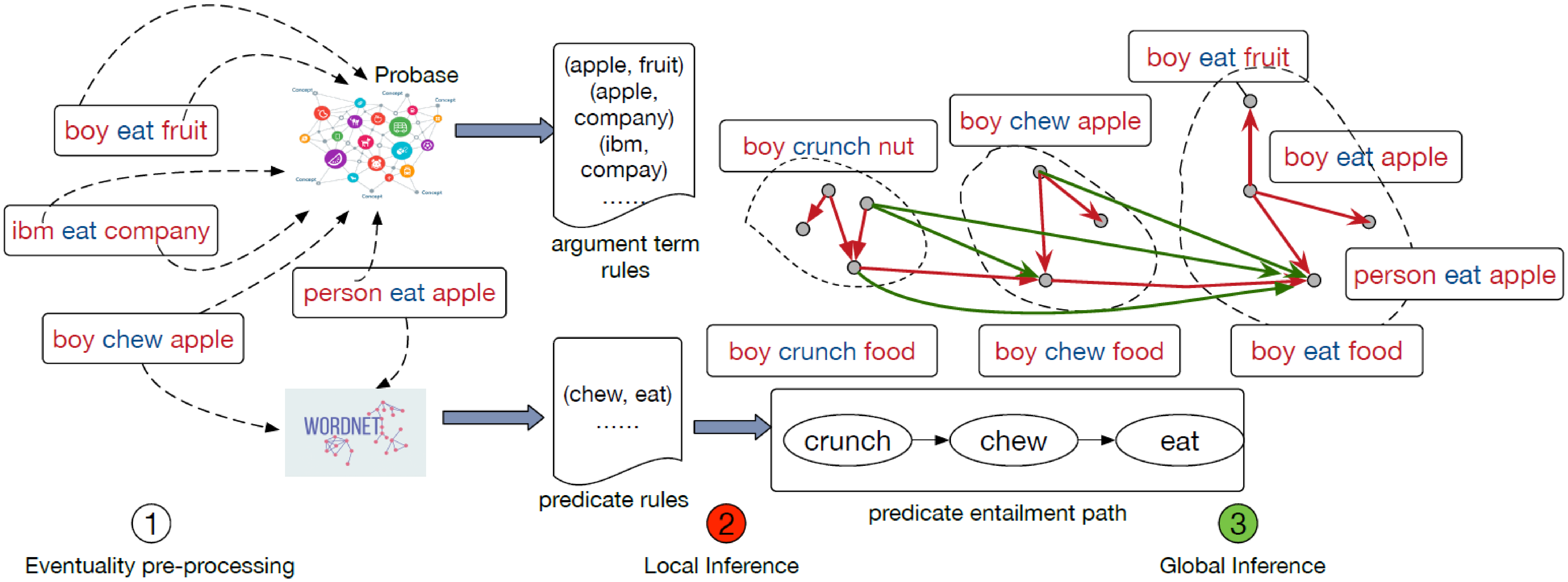


Entailment Graph Construction



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

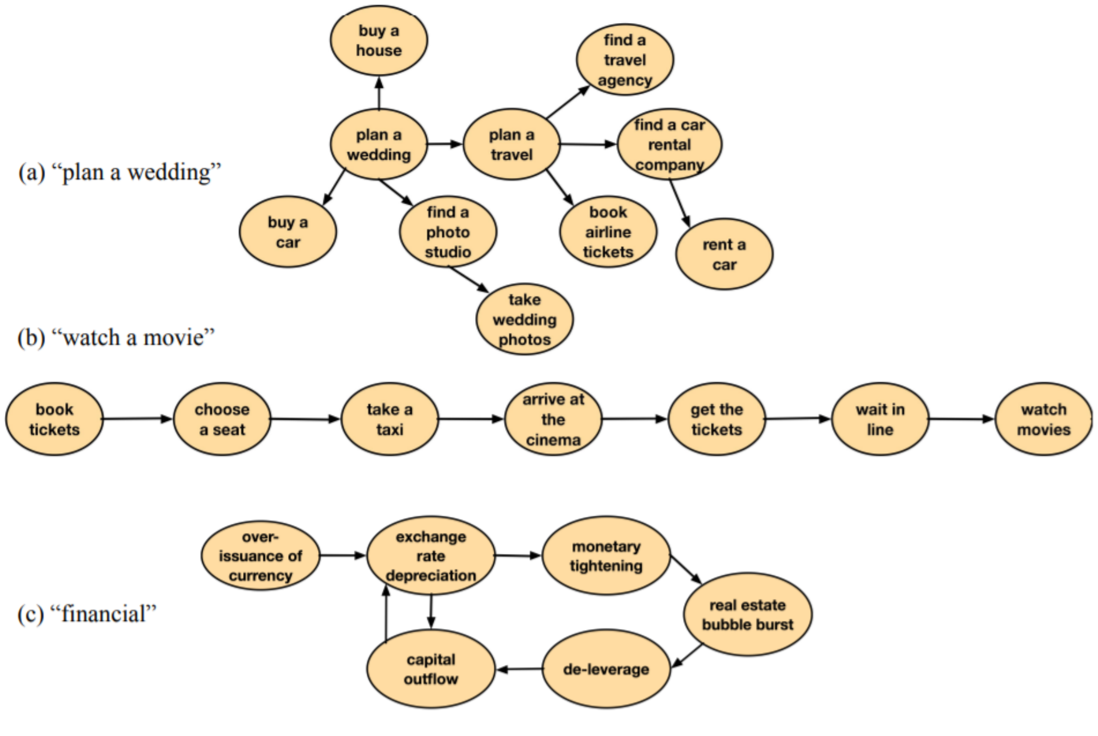
Three-step Construction

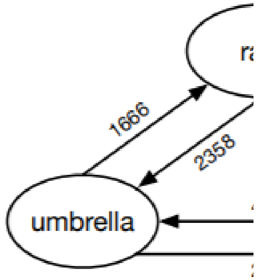


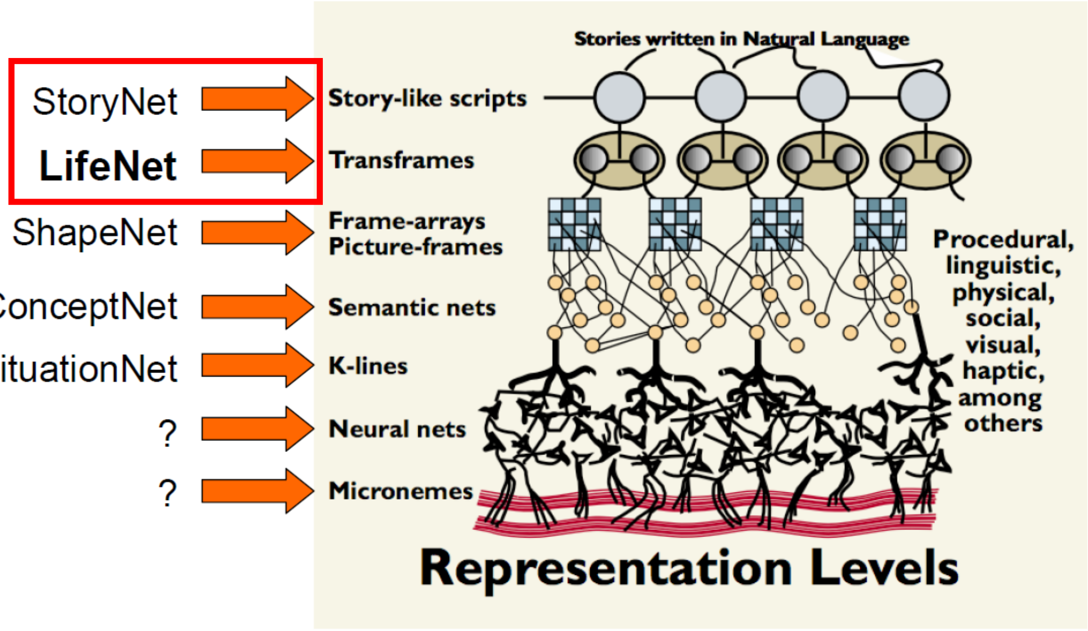
Changlong Yu, Hongming Zhang, Yangqiu Song, Wilfred Ng, Lifeng Shang . Enriching Large-Scale Eventuality Knowledge Graph with Entailment Relations. AKBC. 2020.

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







ELG: An Event Logic Graph Xiao Ding, Zhongyang Li, Ting Liu, Kuo Liao. Arxiv. 2019.
Zhongyang Li, Xiao Ding, Ting Liu, J Edward Hu, and Benjamin Van Durme. Guided generation of cause and eff

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)

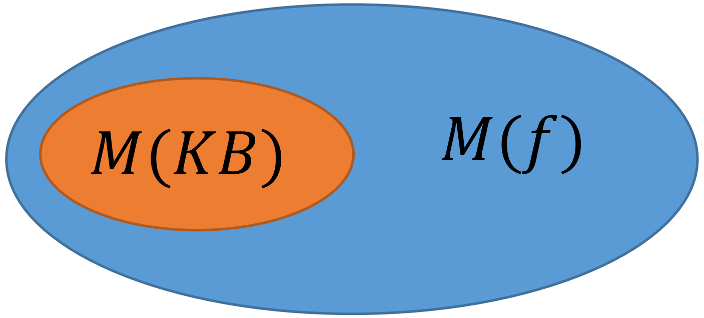
Learning and Reasoning with CSKB/CSKG

- Introduction
- Learning and Reasoning on CSKBs/CSKGs
- Learning and Reasoning for Downstream Tasks (CSQA)

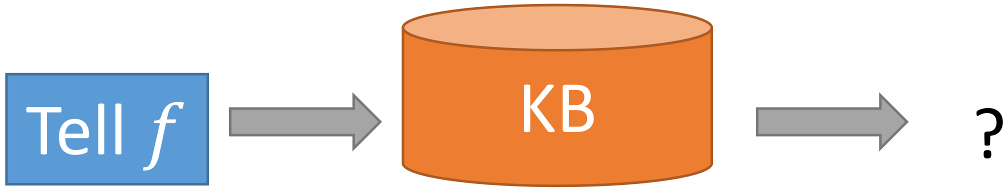
Slides credit of this part: Tianqing Fang

Reasoning

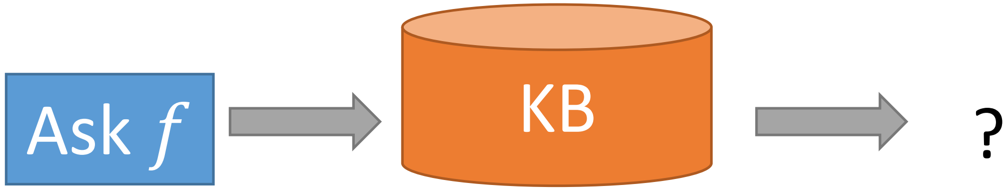
- General reasoning
  - Logical reasoning: Given premise/presumption, draw conclusions based ***solely*** on the premise
- For example
  - $KB = \{Rain \rightarrow Wet, Rain\}, f = Wet$
  - Applying Modus ponens inference rule in KB:
  - $\frac{Rain, Rain \rightarrow Wet}{Wet}$



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



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



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

Commonsense Reasoning

- Commonsense reasoning in natural language:
  - Logical reasoning: E.g., first-order `IsA` relations. Taxonomy reasoning. (Davis 2017)
  - General natural language: Draw conclusions similar to humans' *folk psychology* and *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 `IsA`
  - KB may be noisy. Needs probabilistic reasoning
  - Implicit inferential knowledge outside of the taxonomy

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

If X hit Y on the face, Y will be (a) upset (b) happy

Davis, Ernest (25 August 2017). "Logical Formalizations of Commonsense Reasoning: A Survey". Journal of Artificial Intelligence Research. 59: 651–723.

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

Induction

X: Item does not fit in container, REASON, item is big

Deduction

Trophy is an item; Suitcase is a container

Conceptualization

Instantiation

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

Grounding

Consistent

CSKB/Training Data

- Computer not fit in parcel, REASON, Computer is big
- Rock not fit in carrier, REASON, rock is big
- …

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.

If we instantiate all, it's possible to entail

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)

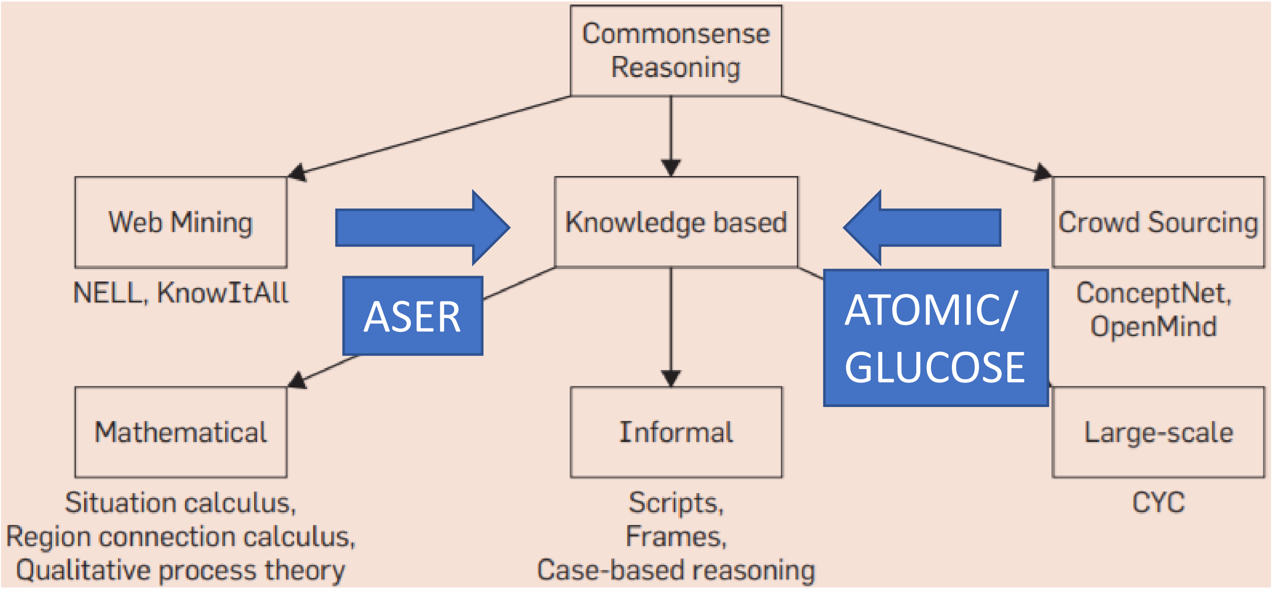
Inference with Entailment

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

- 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

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. SemEval@NAACL-HLT, 2012.

Reasoning Approaches and Typical Objectives (2015)



	Math-based	Informal	Large-scale	Web mining	Crowd sourcing
Architecture	Substantial	Little	Substantial	Moderate	Little
Plausible reasoning	Substantial	Moderate	Substantial	Little	Little
Range of reasoning modes	Moderate	Substantial	Moderate	Little	Little
Painstaking fundamentals	Substantial	Little	Moderate	Little	Little
Breadth	Little	Moderate	Substantial	Substantial	Substantial
Independence of experts	Little	Little	Little	Substantial	Substantial
Concern with applications	Moderate	Substantial	Substantial	Moderate	Moderate
Cognitive modeling	Little	Substantial	Little	Little	Moderate

- **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.

Davis, Ernest, and Gary Marcus. "Commonsense reasoning and commonsense knowledge in artificial intelligence. " *Communications of the ACM* 58.9 (2015): 92-103.

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

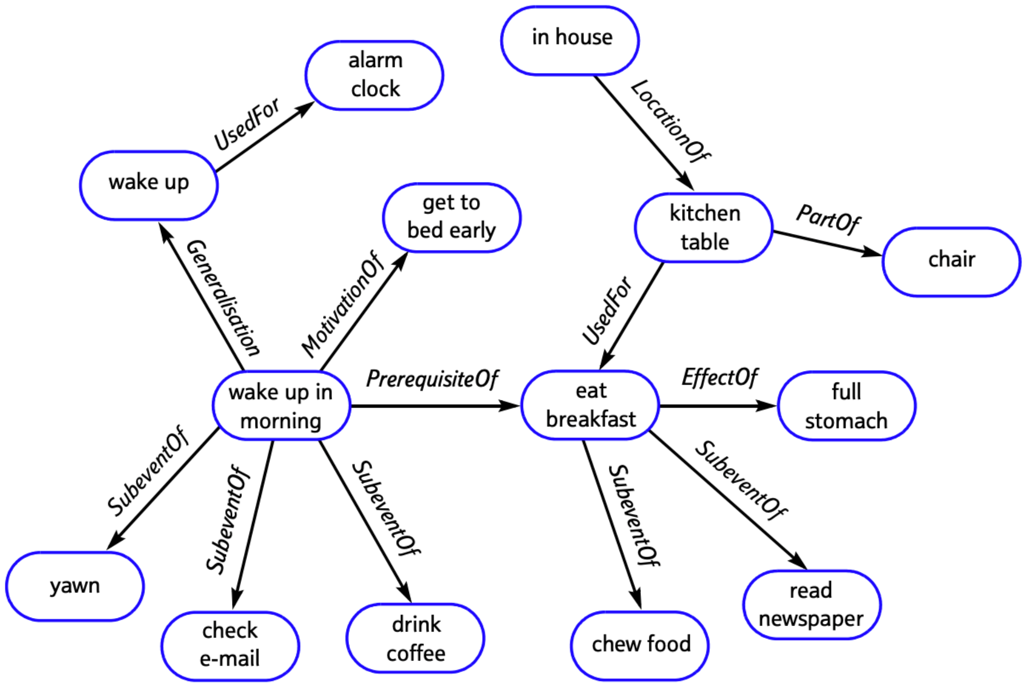
Commonsense Resources and Benchmarks

- 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

Dreifus C: ‘Got stuck for a moment: an interview with Marvin Minsky’, International Herald Tribune (August 1998).

Commonsense Knowledge Bases

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

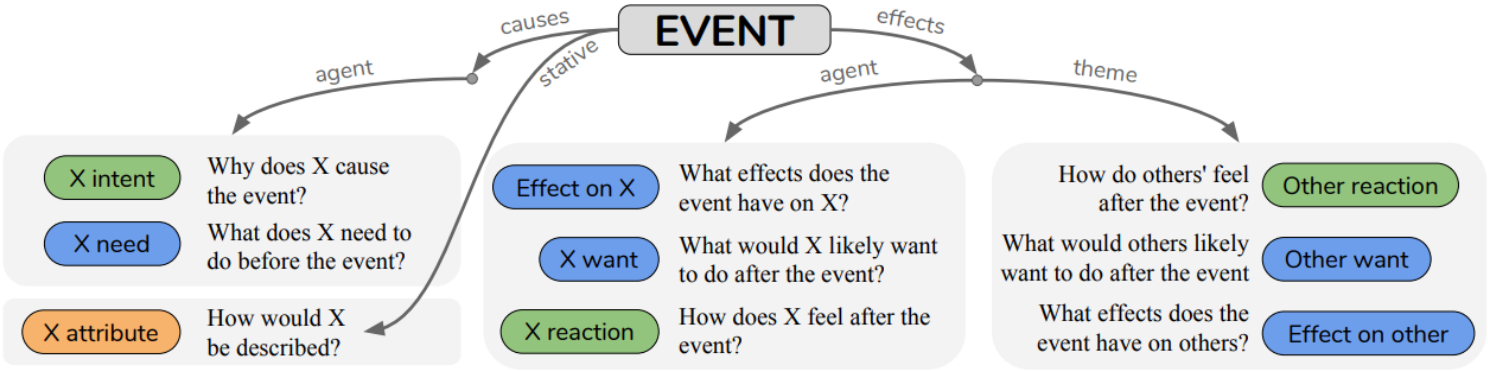


Speer, Robyn, Joshua Chin, and Catherine Havasi. "Conceptnet 5.5: An open multilingual graph of general knowledge." *AAAI*. 2017.

Commonsense Knowledge Bases

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

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



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

SomeoneA possesses Something
Enables
SomeoneA moves it

Sap, Maarten, et al. "Atomic: An atlas of machine commonsense for if-then reasoning." *AAAI 2019.*
Mostafazadeh, Nasrin, et al. "GLUCOSE: GeneraLized and COntextualized Story Explanations." *EMNLP 2020*

Commonsense Resources and Benchmarks

- Scale and Comparisons of Large-scale CSKBs

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

Learning and Reasoning with CSKB/CSKG

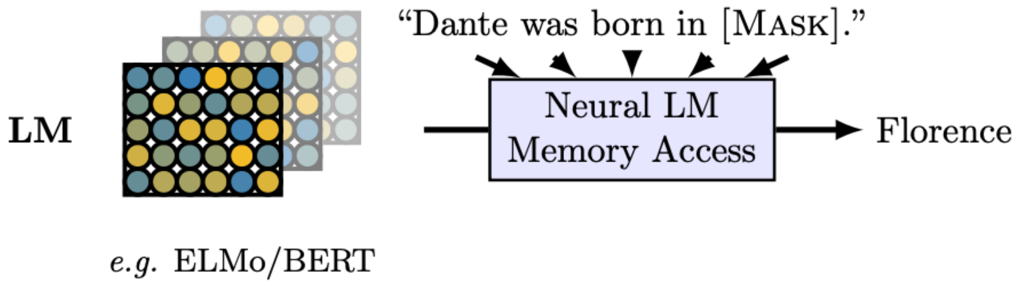
- Introduction

- 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

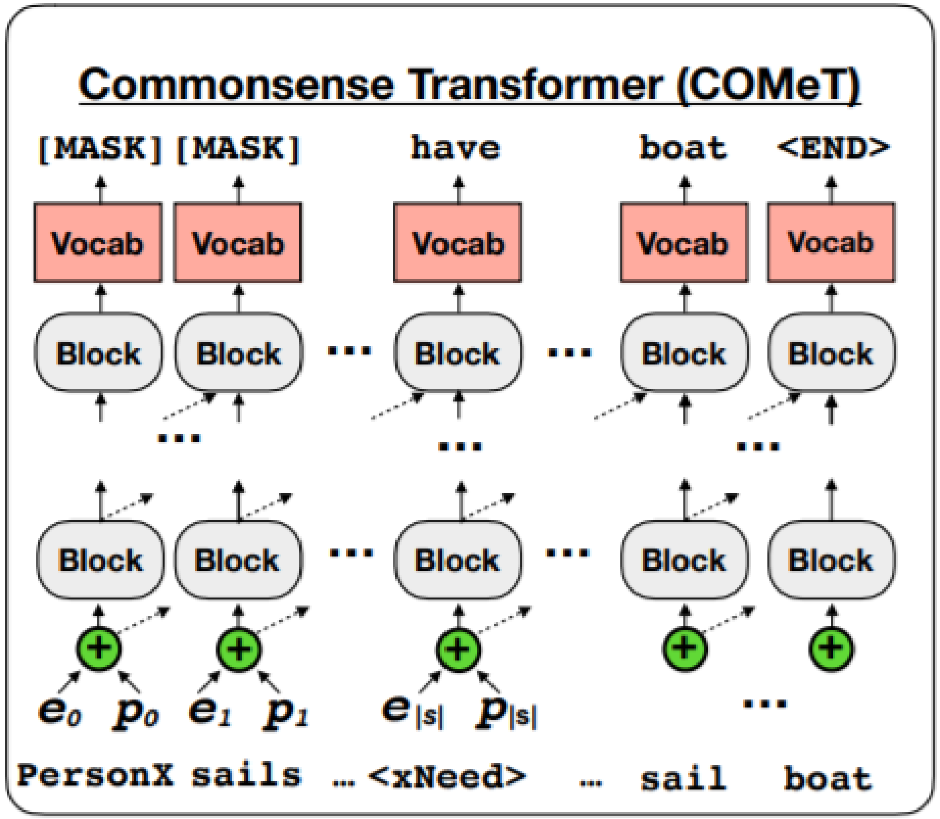


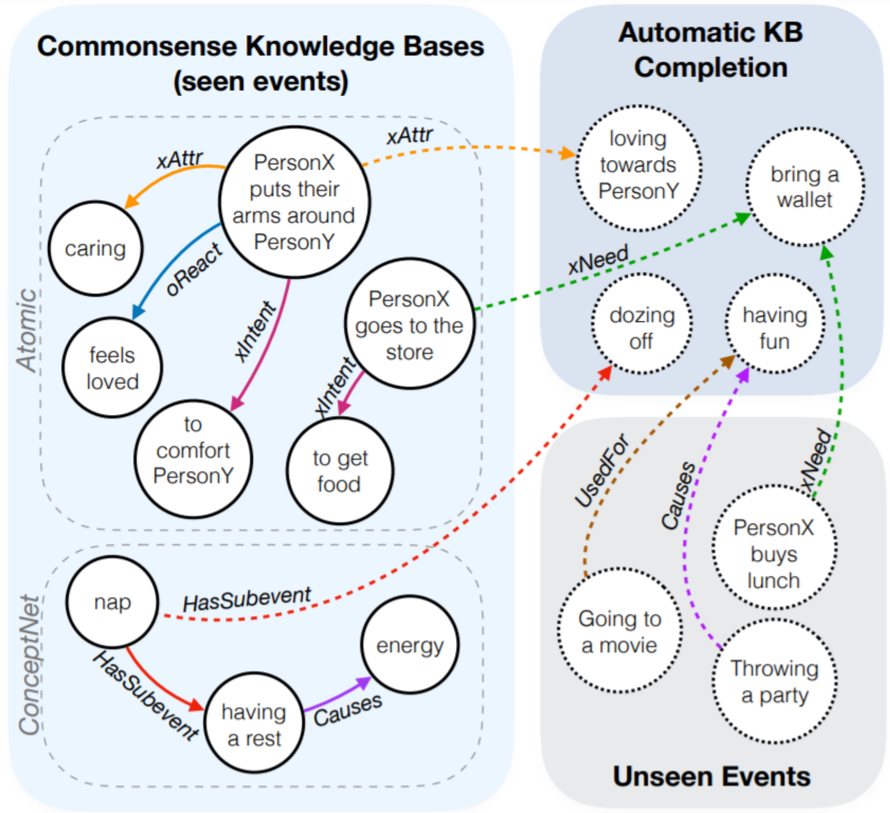
- Cloze style
  - LAMA
    - English ConceptNet, single-token objects.
    - $(Head, Relation, [MASK])$
  - Mining ConceptNet knowledge using PTLM
    - Turning triples to sentences
      - (ferret, `AtLocation`, pet store) -> ferret is in the pet store
    - Generate tails using GPT and BERT
  - A lot of prompt-based methods have been developed

Petroni, Fabio, et al. "Language Models as Knowledge Bases?." EMNLP 2019
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





COMET: Commonsense Transformers for Automatic Knowledge Graph Construction Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, Yejin Choi. ACL, 2019.

Symbolic Knowledge Distillation

Corpus	Accept	Reject	N/A	Size	Size (div)
ATOMIC^{20}_{20}	86.8	11.3	1.9	0.6M	0.56
$\text{ATOMIC}^{\mathbf{10x}}$	78.5	18.7	2.8	**6.5M**	**4.38**
	88.4	9.5	2.1	5.1M	3.68
	91.5	6.8	1.7	4.4M	3.25
	94.3	4.6	1.1	3.6M	2.74
	95.3	3.8	1.0	3.0M	2.33
	96.4	2.7	0.8	2.5M	2.00

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

CKG Completion Model	Train Corpus Acc	Accept	Reject	N/A
GPT2-XL zero-shot	-	45.1	50.3	4.6
GPT-3	-	73.3	24.1	2.6
COMET^{20}_{20}	86.8	81.5	16.3	2.2
$\text{COMET}^{\text{DIS}}_{\text{TIL}}$	78.5	78.4	19.2	2.4
$+\text{critic}_{\text{low}}$	91.5	82.9	14.9	2.2
$+\text{critic}_{\text{high}}$	96.4	**87.5**	10.2	2.3

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

```
What needs to be true for this
event to take place?

...

Event <i>:  X goes jogging

Prerequisites:  For this to
happen, X needed to wear running
shoes

...

Event <i>:  X looks at flowers

Prerequisites:  For this to
happen,
```

- A set of 100 high-quality events from ATOMIC^{20}_{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

Symbolic Knowledge Distillation: from General Language Models to Commonsense Models Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, Yejin Choi. 2021.

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

	#Nodes	#Edges	Avg In-Degree
ConceptNet	78,088	10,000	1.25
ATOMIC	256,570	610,536	2.25
FB15K-237	14,505	272,115	16.98

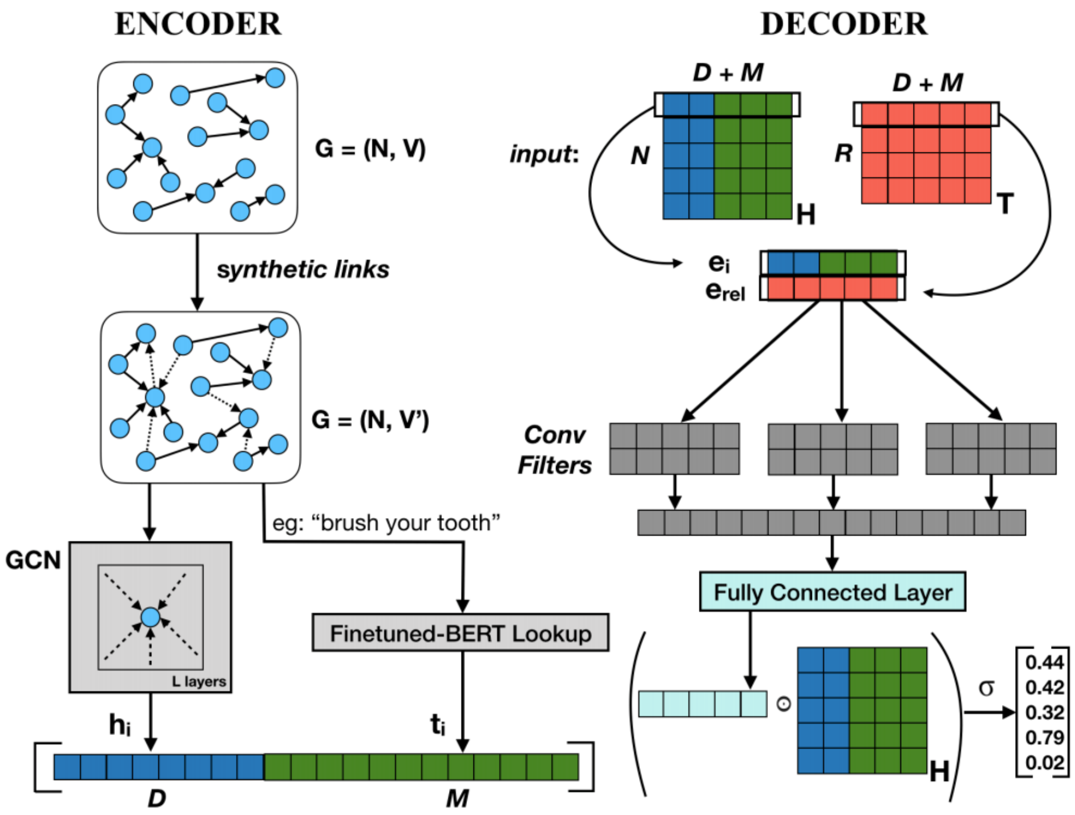
  - Need to deal with sparsity in CSKB.
  - Need to encode semantics of the nodes.

Malaviya, Chaitanya, et al. “Commonsense knowledge base completion with structural and semantic context.” AAAI 2020.

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

	CN-100K				ATOMIC			
	MRR	HITS@1	@3	@10	MRR	HITS@1	@3	@10
DISTMULT	8.97	4.51	9.76	17.44	12.39	9.24	15.18	18.30
COMPLEX	11.40	7.42	12.45	19.01	**14.24**	**13.27**	14.13	15.96
CONVE	20.88	13.97	22.91	34.02	10.07	8.24	10.29	13.37
CONVTRANSE	18.68	7.87	23.87	38.95	12.94	12.92	12.95	12.98
COMET-NORMALIZED	6.07	0.08	2.92	21.17	3.36*	0.00*	2.15*	15.75*
COMET-TOTAL	6.21	0.00	0.00	24.00	4.91*	0.00*	2.40*	21.60*
BERT + CONVTRANSE	49.56	38.12	55.5	71.54	12.33	10.21	12.78	16.20
GCN + CONVTRANSE	29.80	21.25	33.04	47.50	13.12	10.70	13.74	17.68
SIM + GCN + CONVTRANSE	30.03	21.33	33.46	46.75	13.88	11.50	**14.44**	**18.38**
GCN + BERT + CONVTRANSE	50.38	38.79	56.46	72.96	10.8	9.04	11.21	14.10
SIM + GCN + BERT + CONVTRANSE	**51.11**	**39.42**	**59.58**	**73.59**	10.33	8.41	10.79	13.86



$$s(\mathbf{e}_i, \mathbf{rel}, \mathbf{e}_j) = \sigma(M(e_i, e_{rel})W_{conv}e_j)$$

Malaviya, Chaitanya, et al. "Commonsense knowledge base completion with structural and semantic context." AAAI 2020.

InductivE

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

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

Model	CN-100K			CN-82K			ATOMIC		
	MRR	Hits@3	Hits@10	MRR	Hits@3	Hits@10	MRR	Hits@3	Hits@10
DistMult	10.62	10.94	22.54	2.80	2.90	5.60	12.39	**15.18**	18.30
ComplEx	11.52	12.40	20.31	2.60	2.70	5.00	**14.24**	14.13	15.96
ConvE	20.88	22.91	34.02	8.01	8.67	13.13	10.07	10.29	13.37
RotatE	24.72	28.20	45.41	5.71	6.00	11.02	11.16	11.54	15.60
COMET	6.07	2.92	21.17	-	-	-	4.91	2.40	**21.60**
Malaviya et al.	52.25	58.46	73.50	16.26	17.95	27.51	13.88	14.44	18.38
InductivE	**57.35**	**64.50**	**78.00**	**20.35**	**22.65**	**33.86**	14.21	14.82	20.57
Improvement	9.8%	10.3%	6.1%	25.2%	26.2%	23.1%	2.38%	2.63%	11.92%

Wang, Bin, et al. "Inductive Learning on Commonsense Knowledge Graph Completion." IJCNN, 2021.

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

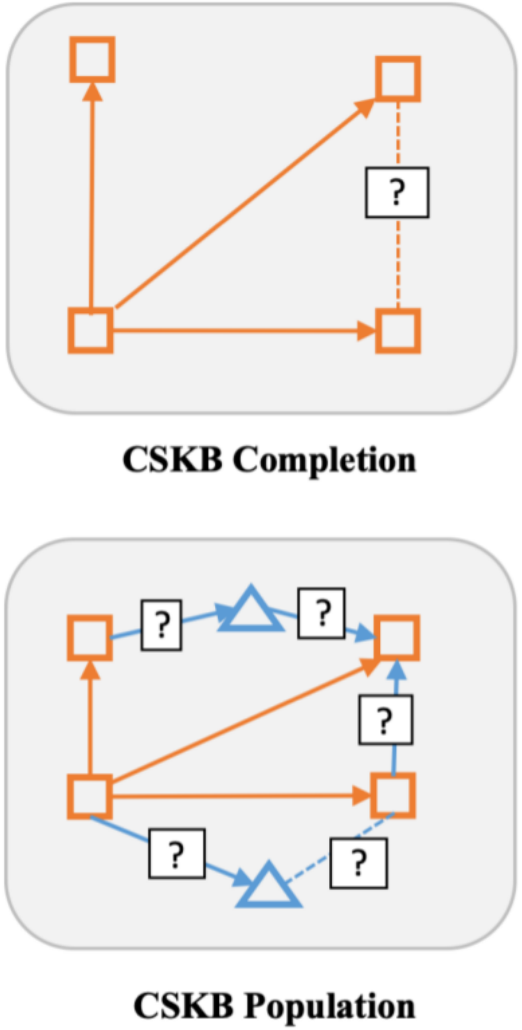
- Denote the CSKB as $\mathcal{C} = \{(h, r, t) | h \in \mathcal{H}\ , r \in \Re, t \in \mathcal{T}\}$. An automatically extracted eventuality knowledge graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which is much larger than $\mathcal{C}$.

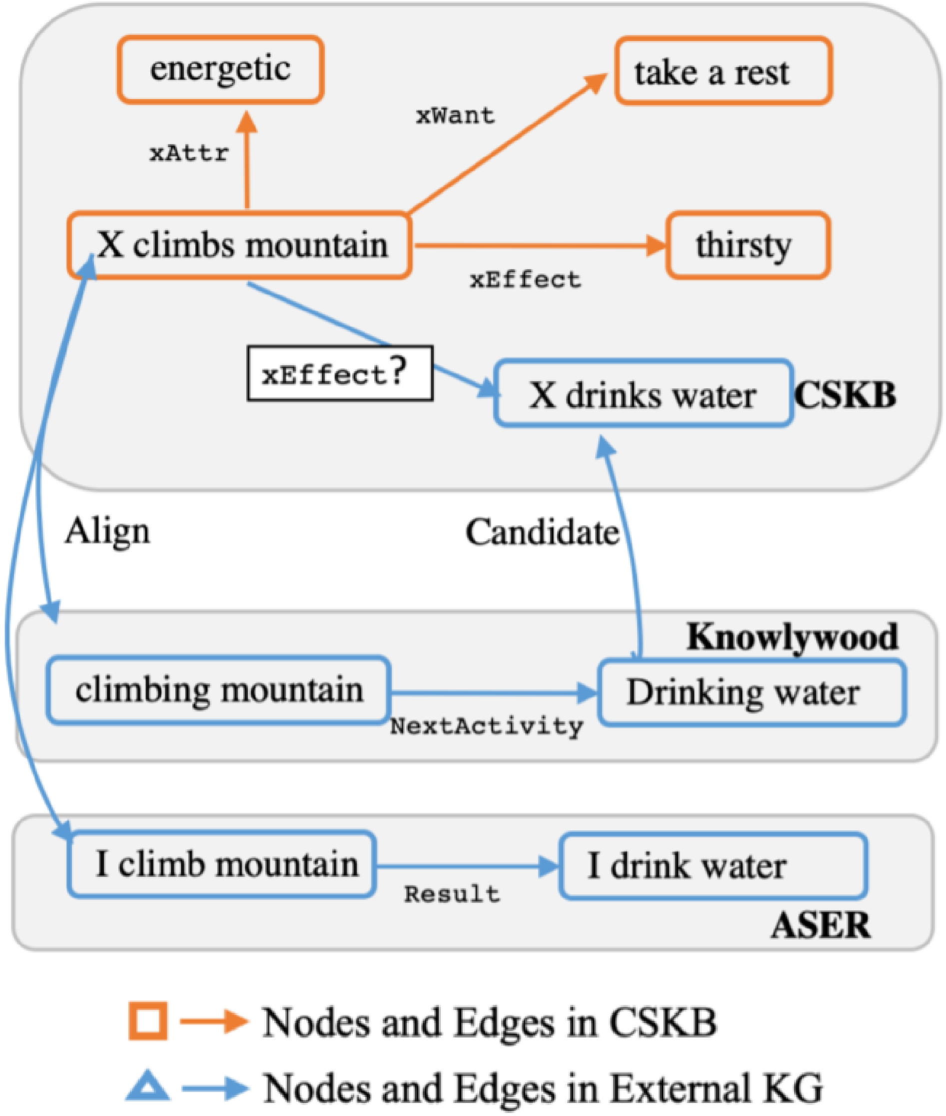
- Denote $\mathcal{G}^{\mathcal{C}}$ as the graph by aligning $\mathcal{G}$ and $\mathcal{C}$.

- 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

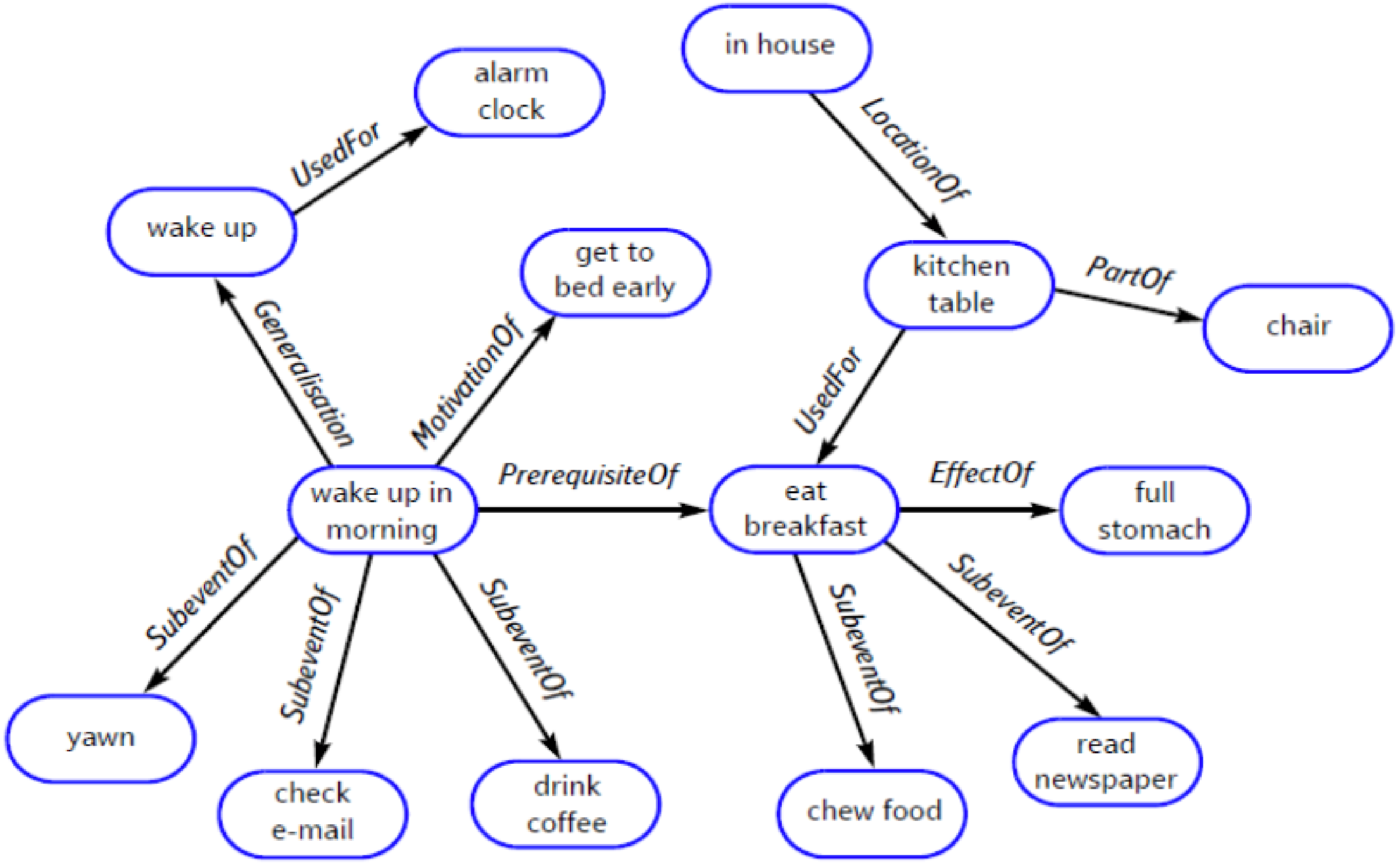
- ATOMIC Population
  - Latent variables (events and states) of commonsense

Slides credit for this part: Hongming Zhang

ConceptNet (Speer & Havasi, 2012)

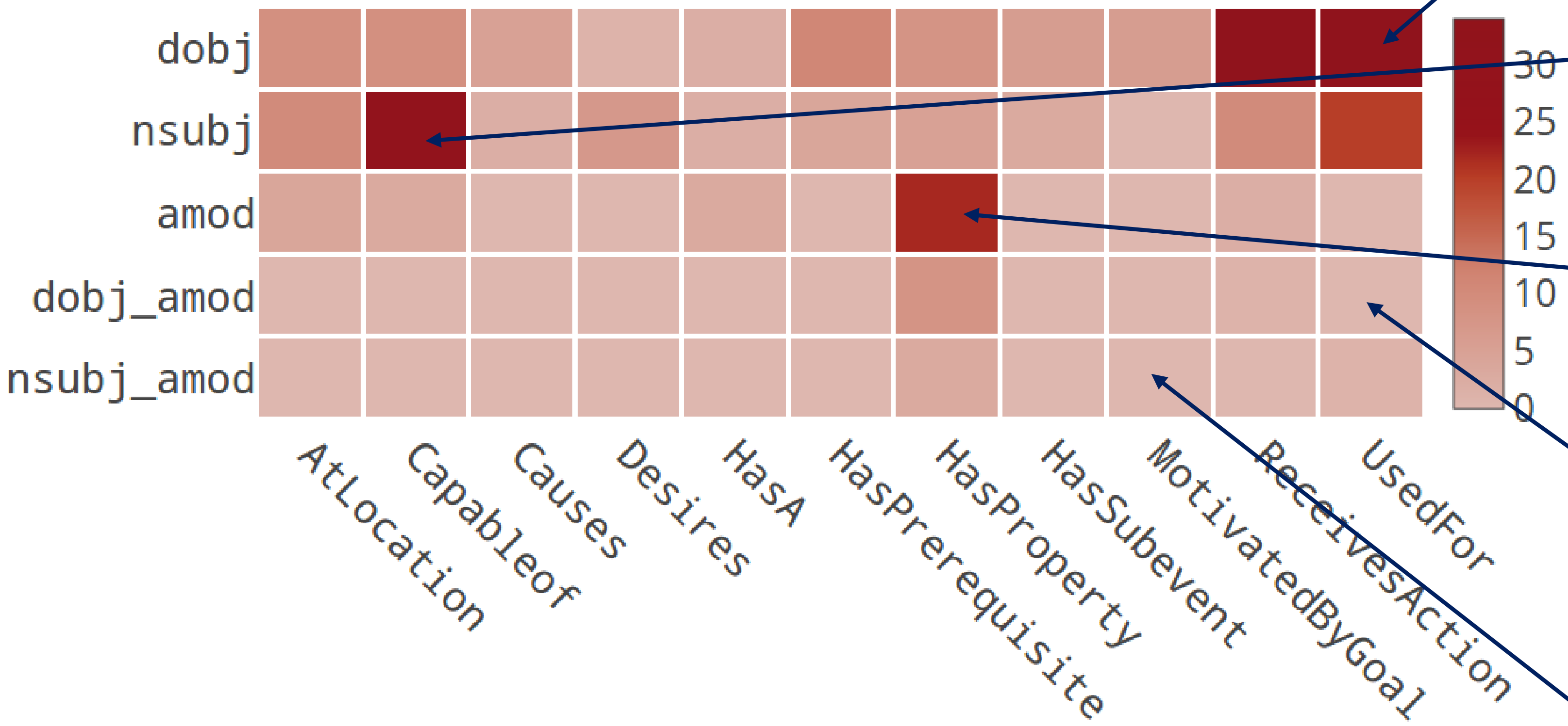
Core is OMCS (Liu & Singh 2004)

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



Speer and Havasi. "Representing General Relational Knowledge in ConceptNet 5." *LREC*. 2012.

Revisit the Correlations of Selectional Preference and OMCS



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

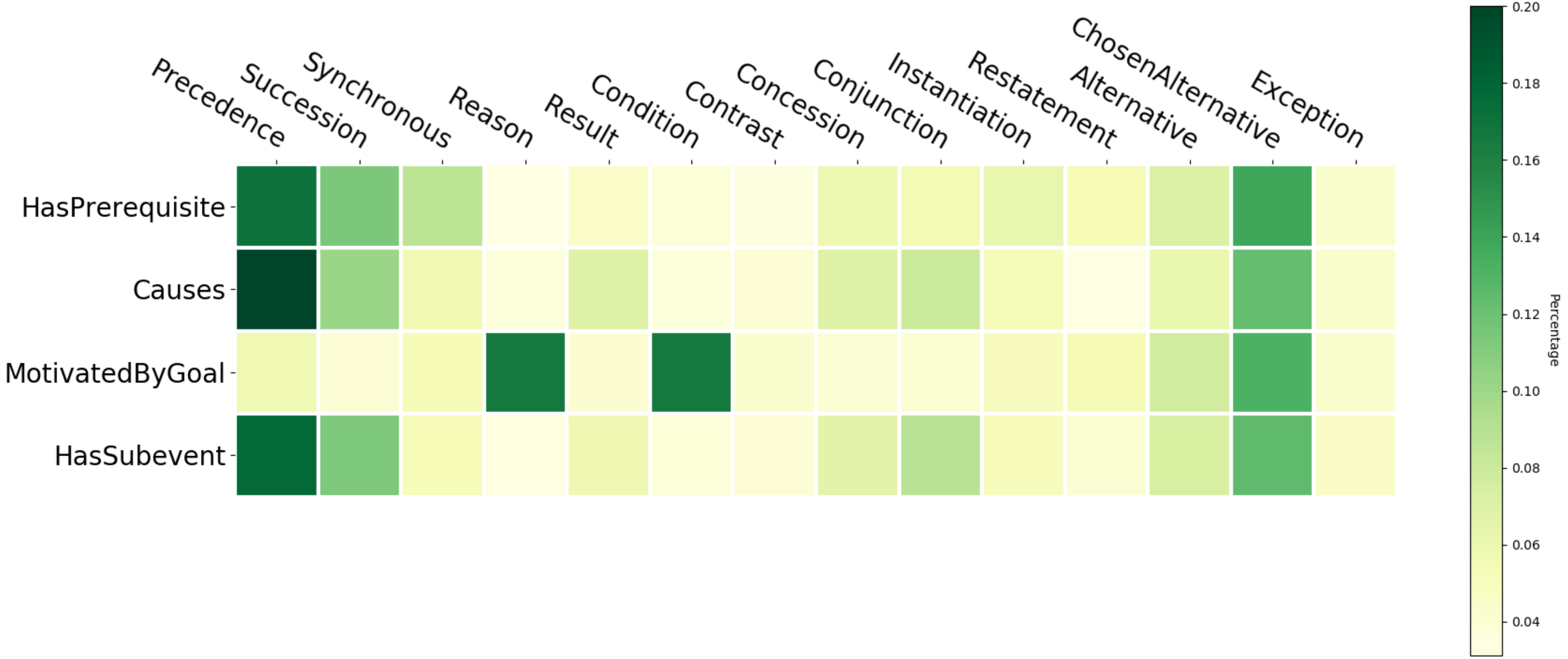
(phone, ring) (nsubj, 8.75)
(phone, CapableOf, ring)

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

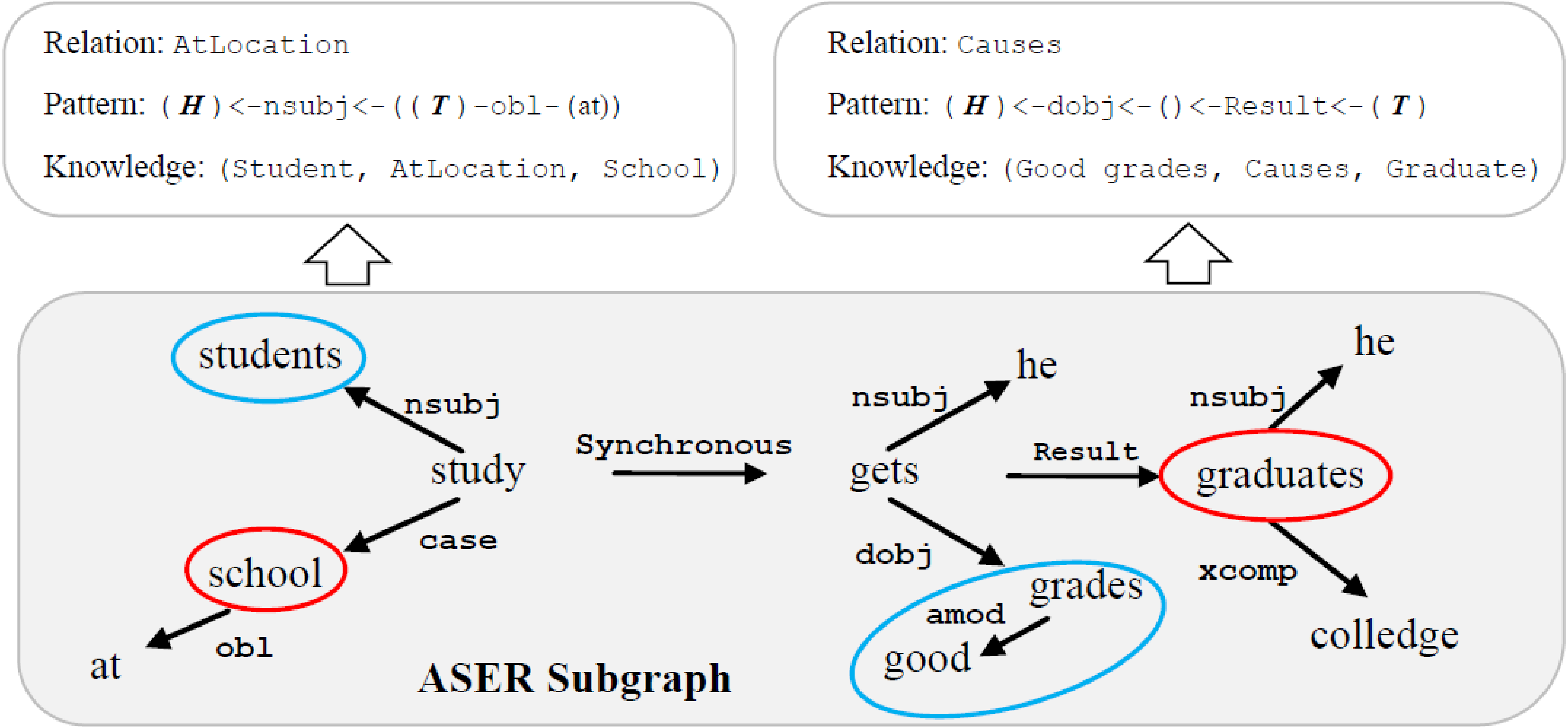
(create, new) (dobj_amod, 8.25)
(create idea, UsedFor, invent new things)

(hungry, eat) (nsubj_amod, 10.00)
(eat, MotivatedByGoal, are hungry)

Revisit the Correlations of ASER and OMCS



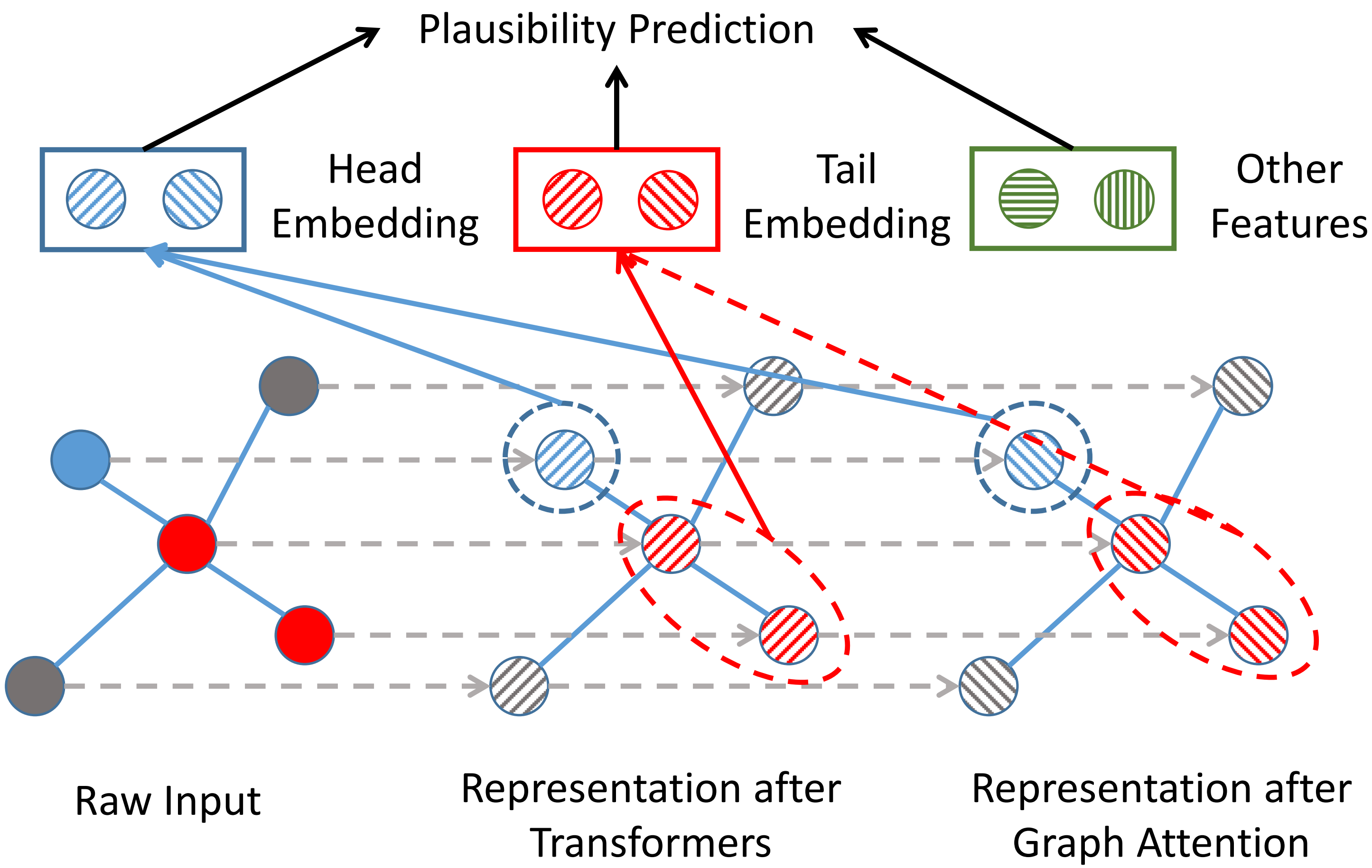
TransOMCS



Relation: AtLocation
Pattern: (H)<-nsubj<-((T)-obl-(at))
Knowledge: (Student, AtLocation, School)
Relation: Causes
Pattern: (H)<-dobj<-()<-Result<-(T)
Knowledge: (Good grades, Causes, Graduate)
students
nsubj
study
case
school
obl
at
Synchronous
ASER Subgraph
he
nsubj
gets
dobj
grades
amod
good
Result
graduates
he
nsubj
xcomp
colledge

Knowledge Ranking

- Assigning confidence score to each piece of extracted commonsense
  - Leverage the semantics of the original sentences
  - Leverage the frequency information



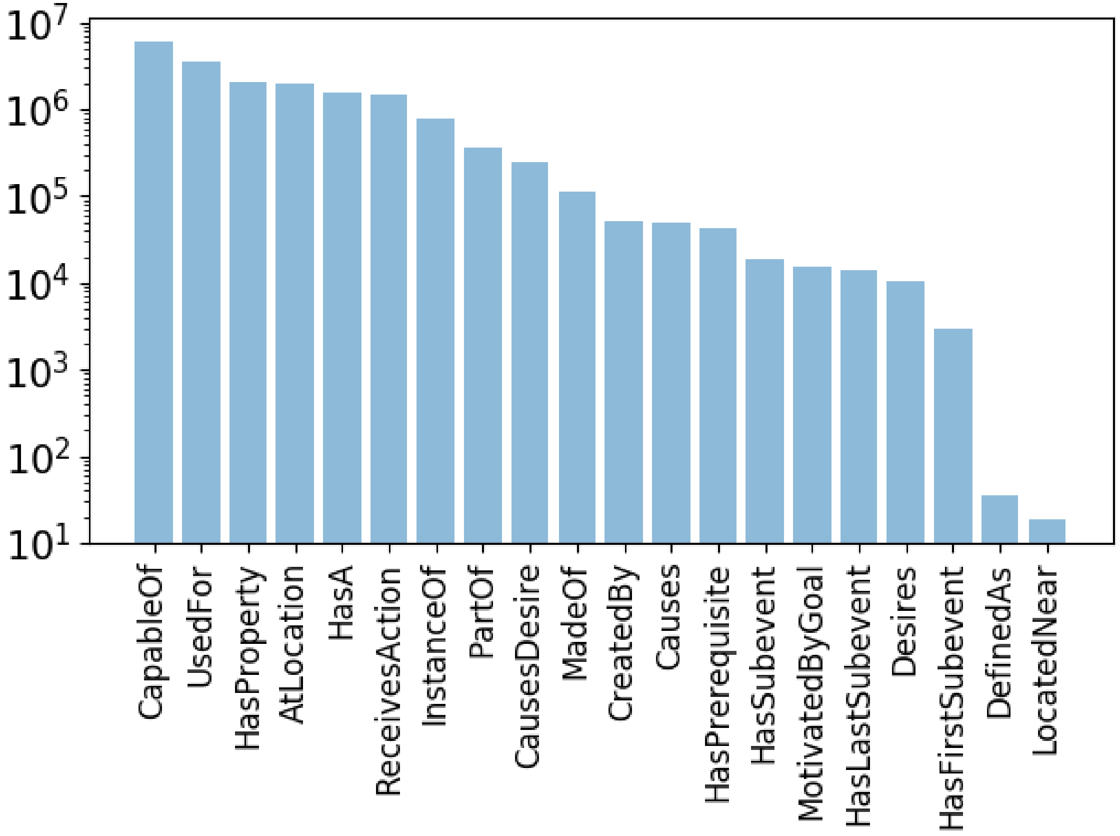
Transferring ASER to ConceptNet

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

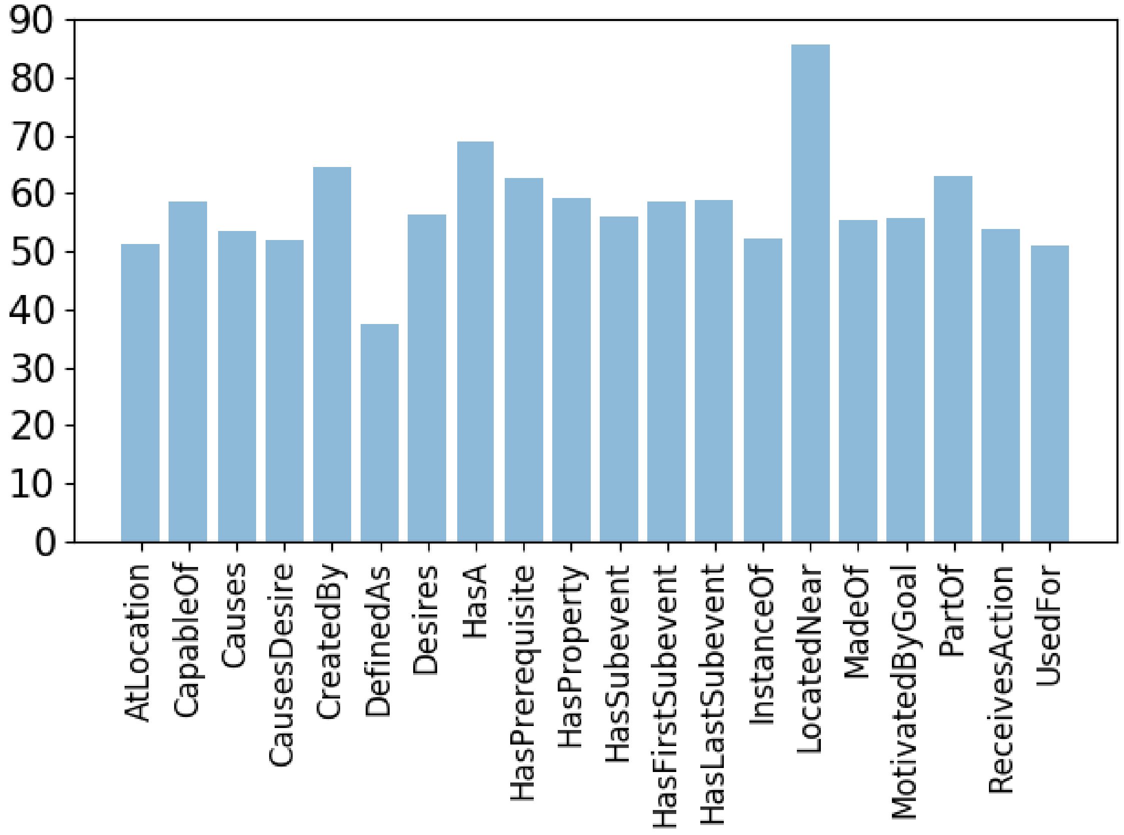
Transferability from linguistic knowledge to commonsense knowledge

SP over eventualities can effectively represent interesting commonsense knowledge

Distribution of Relations and Accuracy



Distribution of Relations



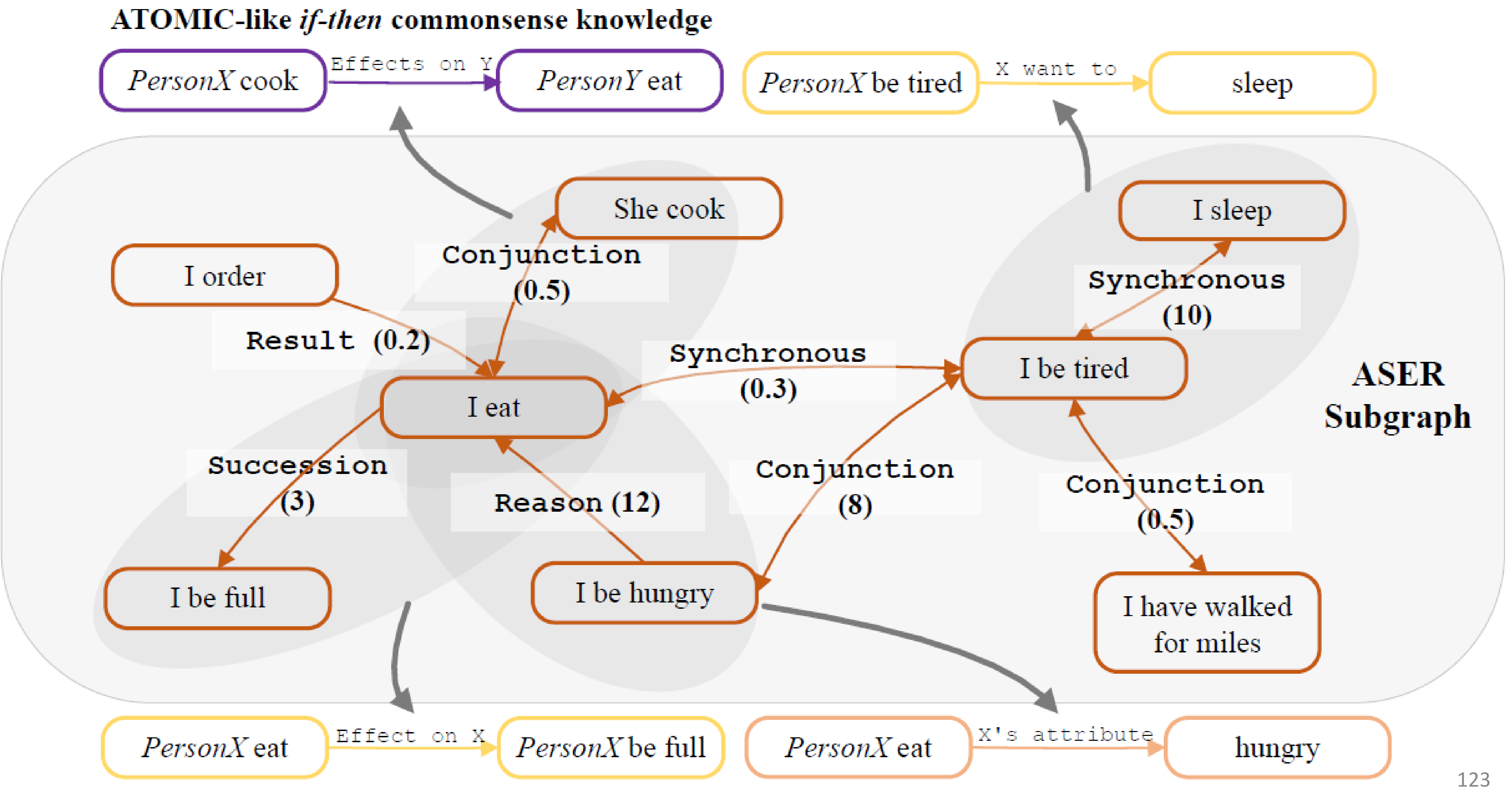
Accuracy

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

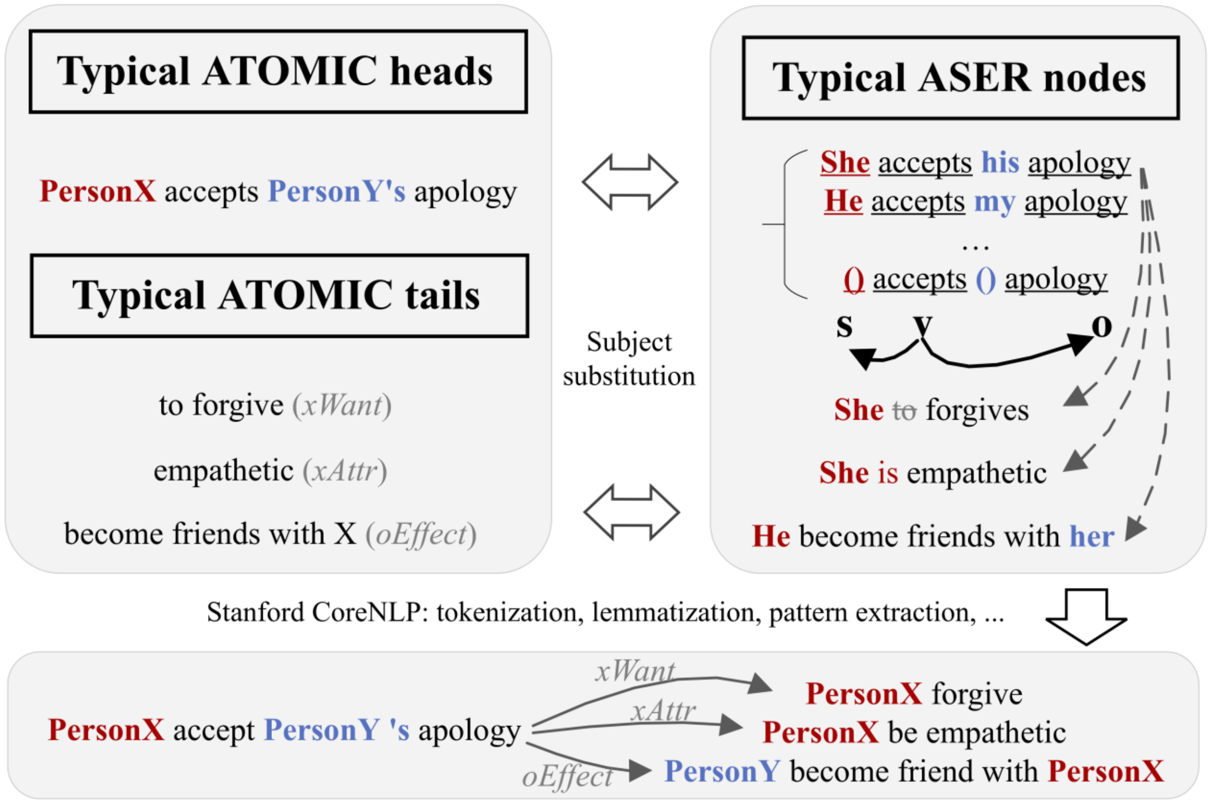
	$ASER_{norm}$ Coverage				Avg. Degree in $ASER_{norm}$				Avg. Degree in $\mathcal{C}$			
					In-Degree		Out-Degree		In-Degree		Out-Degree	
	head(%)	tail(%)	edge(%)	#hops	head	tail	head	tail	head	tail	head	tail
ATOMIC	79.76	77.11	59.32	2.57	90.9	61.3	91.2	61.6	4.2	3.4	34.6	1.5
$ATOMIC_{20}^{20}$	80.39	47.33	36.73	2.65	96.9	66.9	97.3	67.3	4.3	2.9	34.6	1.5
ConceptNet	77.72	54.79	43.51	2.37	210.7	88.9	211.6	88.9	15.1	8.0	26.2	4.1
GLUCOSE	91.48	91.85	81.01	2.37	224.9	246.4	226.6	248.0	7.2	7.7	6.7	5.5

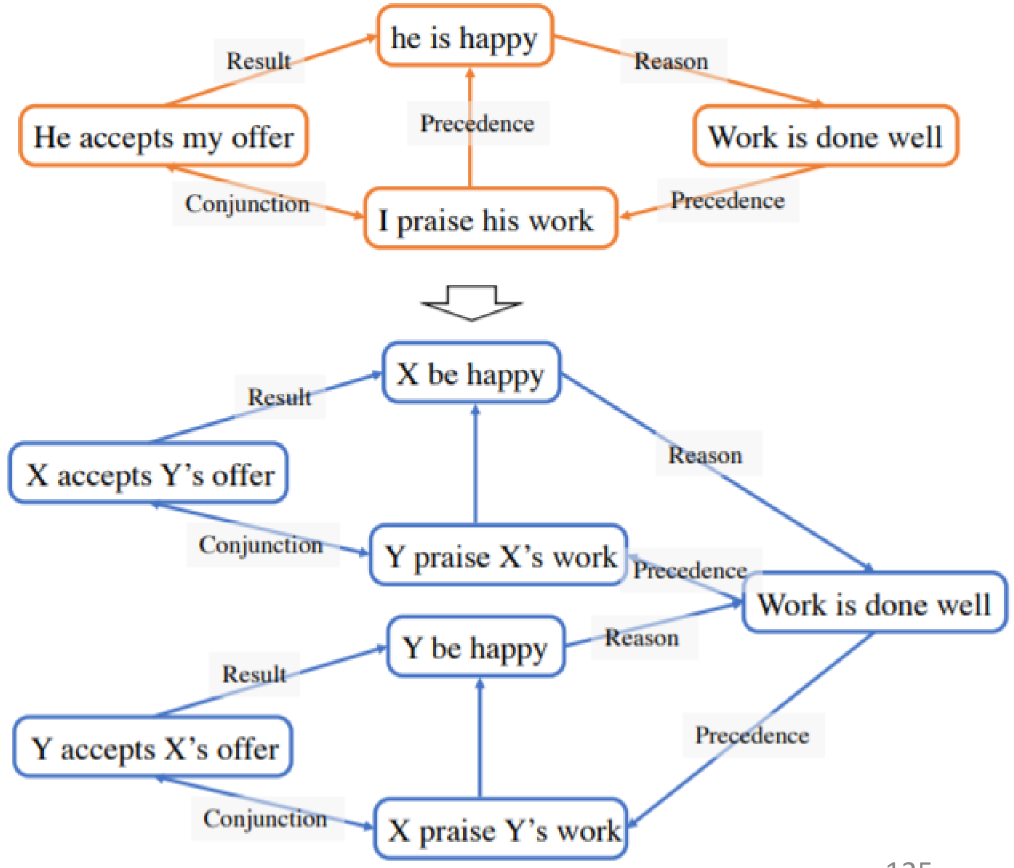
Table 3: The overall matching statistics for the four CSKBs. The *edge* column indicates the proportion of edges where their heads and tails can be connected by paths in ASER. Average (in and out)-degree on $ASER_{norm}$ and $\mathcal{C}$ for nodes from the CSKBs is also presented. The statistics in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

Maarten Sap, et al. ATOMIC: An atlas of machine commonsense for if-then reasoning. AAAI 2019.
Jena D Hwang, et al. (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. AAAI 2021.
Nasrin Mostafazadeh, et al. Glucose: Generalized and contextualized story explanations. NAACL 2020.

Node Alignment with ASER

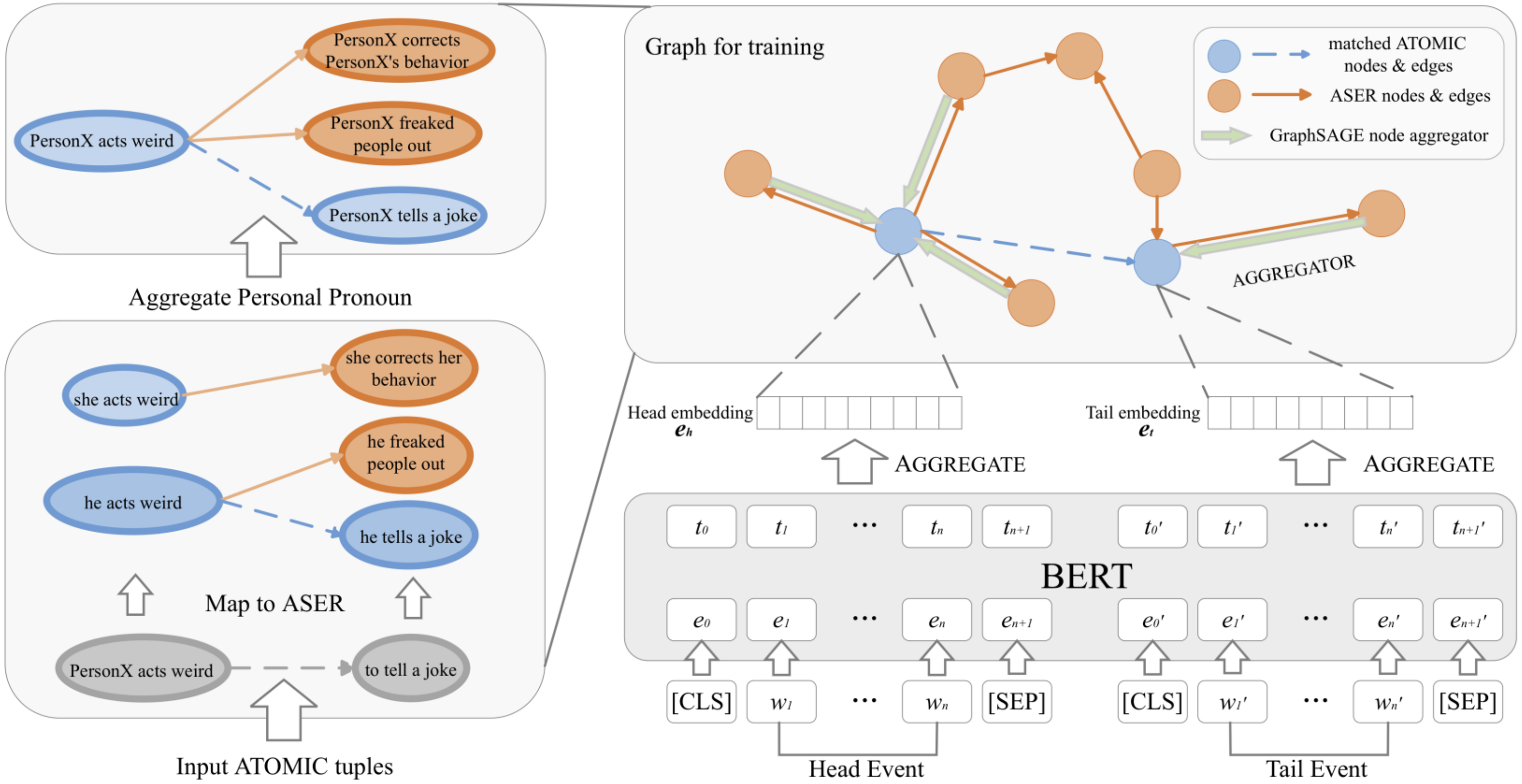
- ASER and other CSKB take different forms of representing personal entities
- Develop simple rules for aligning the two resources.





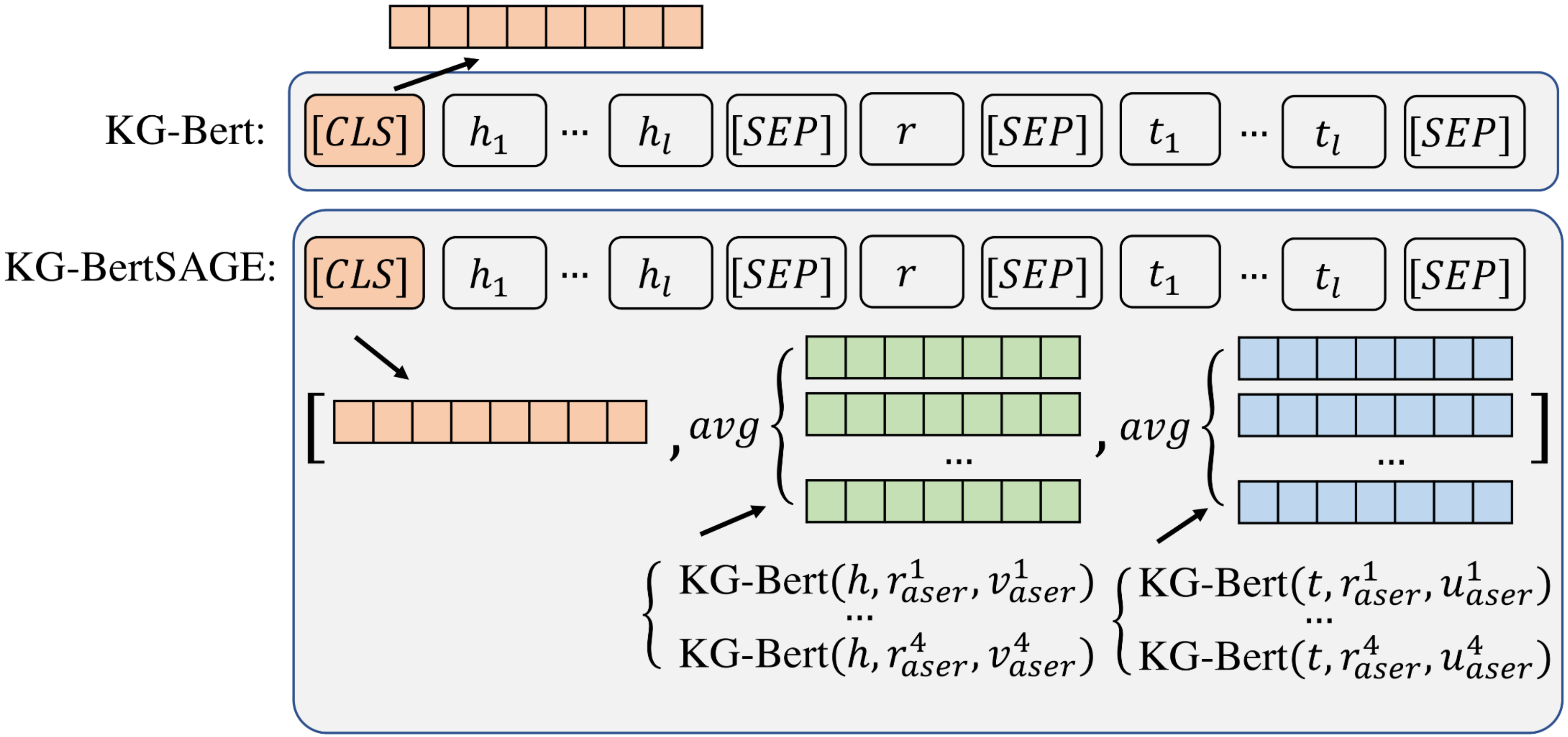
DISCOS (DIScourse to COmmonSense): BertSAGE [WWW 2021]

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



Hamilton, William L., Rex Ying, and Jure Leskovec. "Inductive representation learning on large graphs." NeurIPS. 2017.
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]



Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-Bert: Bert for knowledge graph completion. arXiv preprint arXiv:1909.03193.
Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. Benchmarking Commonsense Knowledge Base Population with an Effective Evaluation Dataset. 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

	Dev	Test	Train
# Triples	6,217	25,514	1,100,362
% Plausible	51.05%	51.74%	-
% Novel Nodes	67.40%	70.01%	-

Relation	ATOMIC$\binom{20}{20}$	ConceptNet	GLUCOSE
oEffect	21,497	0	7,595
xEffect	61,021	0	30,596
gEffect	0	0	8,577
oWant	35,477	0	1,766
xWant	83,776	0	11,439
gWant	0	0	5,138
oReact	21,110	0	3,077
xReact	50,535	0	13,203
gReact	0	0	2,683
xAttr	89,337	0	7,664
xNeed	61,487	0	0
xIntent	29,034	0	8,292
isBefore	18,798	0	0
isAfter	18,600	0	0
HinderedBy	87,580	0	0
xReason	189	0	0
Causes	0	42	26,746
HasSubEvent	0	9,934	0
Total	578,252	10,165	126,776

Relation	number of edges
Precedence	4,957,481
Succession	1,783,154
Synchronous	8,317,572
Reason	5,888,968
Result	5,562,565
Condition	8,109,020
Contrast	23,208,195
Concession	1,189,167
Alternative	1,508,729
Conjunction	37,802,734
Restatement	159,667
Instantiation	33,840
ChosenAlternative	91,286
Exception	51,502
Co_Occurrence	124,330,714
Total	222,994,594

Main Population Results

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

Relation	xWnt	oWnt	gWnt	xEfct	oEfct	gEfct	xRct	oRct	gRct	xAttr	xInt	xNeed	Cause	xRsn	isBfr	isAft	Hndr.	HasSubE.	all
BERT	57.7	64.9	66.3	59.1	66.2	60.0	50.6	**68.7**	72.3	56.2	63.9	56.4	48.3	34.5	59.2	58.0	66.1	73.0	59.4
BERTSAGE	54.7	58.9	58.0	58.0	70.0	54.7	52.8	62.4	**76.6**	55.0	61.0	57.1	46.2	45.5	66.7	64.9	69.6	**80.4**	60.0
KG-BERT	63.2	**69.8**	**69.0**	68.0	70.6	61.0	57.0	64.0	73.8	**59.5**	**64.9**	64.6	47.4	**90.9**	78.0	**77.5**	75.9	68.5	66.1
KG-BERTSAGE	**66.0**	68.9	68.6	**68.2**	**70.8**	**62.3**	**60.5**	64.6	74.1	59.1	63.0	**65.4**	**50.0**	76.4	**78.2**	77.4	**77.5**	67.0	**67.2**
Human	86.2	86.8	83.3	85.2	83.9	79.8	81.1	82.6	76.5	82.6	85.6	87.4	80.1	73.7	89.8	89.9	85.3	85.7	84.4

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.

LR	all	*Original Test Set*	*CSKB head + ASER tail*	*ASER edges*
KGBert	67.5	79.2	57.3	52.6
KGBertSAGE	68.5	80.1	58.2	53.5
GPT2-small	70.5	73.3	64.0	63.0
GPT2-medium	71.5	74.7	65.1	65.1
GPT2-large	71.8	75.5	65.4	65.3
COMET(GPT2XL)	70.4	73.1	64.5	63.7
GPT2XL(ZS)	64.7	65.8	60.8	63.1

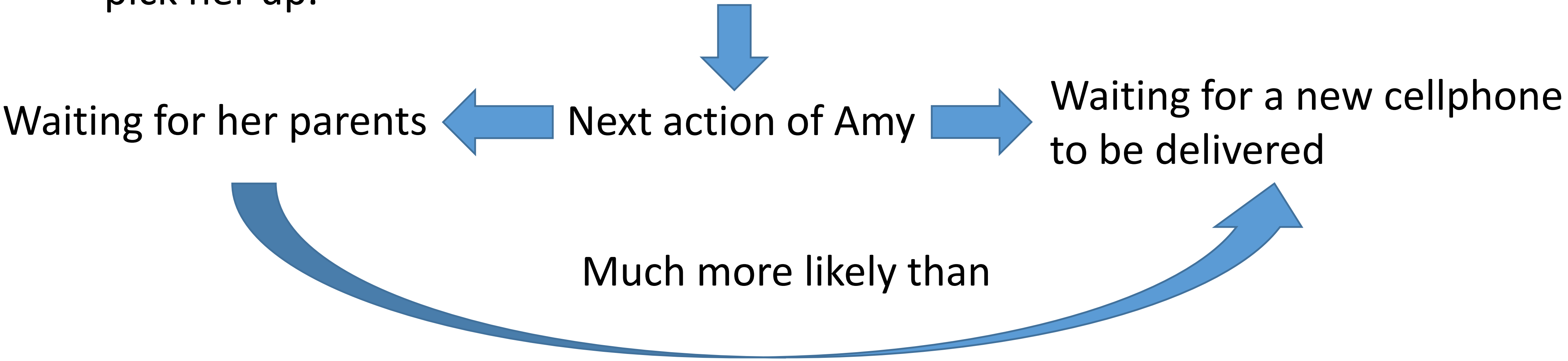
Learning and Reasoning with CSKB/CSKG

- 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.



Liu, Hugo, and Push Singh. "ConceptNet—a practical commonsense reasoning tool-kit." BT technology journal 22.4 (2004): 211-226.

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))

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 7432–7439, 2020.
Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2391–2401, 2019.

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

Liu, Hugo, and Push Singh. "ConceptNet—a practical commonsense reasoning tool-kit." BT technology journal 22.4 (2004): 211-226.
Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. Aser: A large-scale eventuality knowledge graph. In Proceedings ofTheWeb Conference 2020, pages 201–211, 2020.

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).

Next action of Alex

(a) Mop up the floor

(b) Taste the food

(c) Run around in the mess

Much more likely than

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQA: Commonsense reasoning about social interactions. EMNLP-IJCNLP, pages 4453–4463, 2019.

Social Commonsense

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQA: Commonsense reasoning about social interactions. EMNLP-IJCNLP, 2019.
Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A large-scale adversarial dataset for grounded commonsense inference. EMNLP, 2018.

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

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. PIQA: Reasoning about physical commonsense in natural language. AAAI, 2020.

Physical Commonsense

	Sample Question	Sample Answer	Construction Method	Size
PIQA (Bisk et, al., 2020)	How do I find something I lost on the carpet?	(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. ✓	Instructions on everyday events	21K

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. PIQA: Reasoning about physical commonsense in natural language. AAAI, 2020.

Temporal Commonsense

- Commonsense knowledge about time
- Example:
    - taking a vacation

takes longer time than

taking a walk

Temporal Commonsense

	Sample Question	Sample Answer	Construction Method	Size
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?	(1) the aid was denied (2) things started to progress ✓ (3) he received the aid ✓	Human annotations	13K

- 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

unnecessary to be explicitly mentioned in the communication

seven

Numerical Commonsense

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

Category	Example
Objects (35.2%)	A bicycle has *two* tires.
Biology (13.5%)	Ants have *six* legs.
Geometry (11.7%)	A cube has *six* faces.
Unit (6.3%)	There are *seven* days in a week.
Math (7.3%)	…n *nine* now.
Physics (5.7%)	W… …es centigrade.
Geography (2.9%)	…ontinents.
Misc. (17.5%)	…nited States.

Table 1: NUMER… …h category.

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

- There are many other math word problems in NLP

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. Birds have four legs?! NumerSense : Probing numerical commonsense knowledge of pre-trained language models. EMNLP, 2020
Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. NAACL-HLT , 2019.

Spatial Commonsense

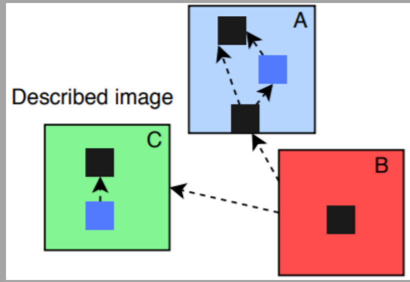
- Cognitive process about spatial objects, relations, and transformations (Clements and Battista, 1992)
- Example:
  - The man is riding a horse (Collell et, al., 2018)

The relative positions of the man and the horse

The man is **above** the horse

Douglas H Clements and Michael T Battista. Geometry and spatial reasoning. Handbook ofresearch on mathematics teaching and learning, pages 420–464, 1992.
Guillem Collell, Luc Van Gool and Marie-Francine Moens. Acquiring common sense spatial knowledge through implicit spatial templates. AAAI 2018.

Spatial Commonsense

	Sample Question	Sample Answer	Construction Method	Size
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

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. SpartQA:: A textual question answering benchmark for spatial reasoning. NAACL 2021.

General Commonsense

- General knowledge involved in everyday situation (e.g., causal commonsense)
- Example:

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

What happened as a RESULT

The liquid in the bottle poured out

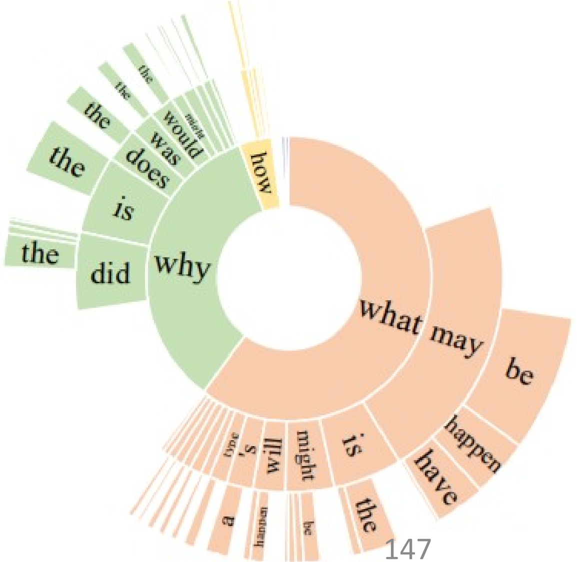
The liquid in the bottle froze

Much more likely than

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. SemEval@NAACL-HLT, 2012.

General Commonsense

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



(a) COSMOS QA

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense k
Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonse
2012.
Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense rea

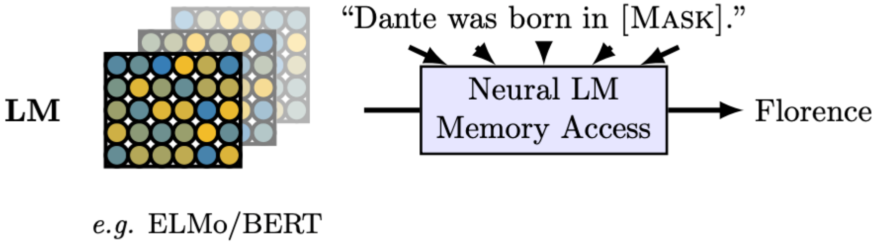
Learning and Reasoning with CSKB/CSKG

- 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 PTLI
  - Fine-tuning(Optional)
  - Prediction



Petroni, Fabio, et al. "Language Models as Knowledge Bases?." EMNLP 2019

Pre-Trained Language Model as the Only Implicit Knowledge Source

- UNICORN (Lourie et, al., 2021)
  - T5-based CSQA model
  - Pre-trained and fined-tuned on a multi-task benchmark – RAINBOW (Lourie et, al., 2021)
  - Sequential training paradigm
  - SOTA on various CSQA benchmarks (e.g., COSMOSQA and PIQA)

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. AAAI, 2021.

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

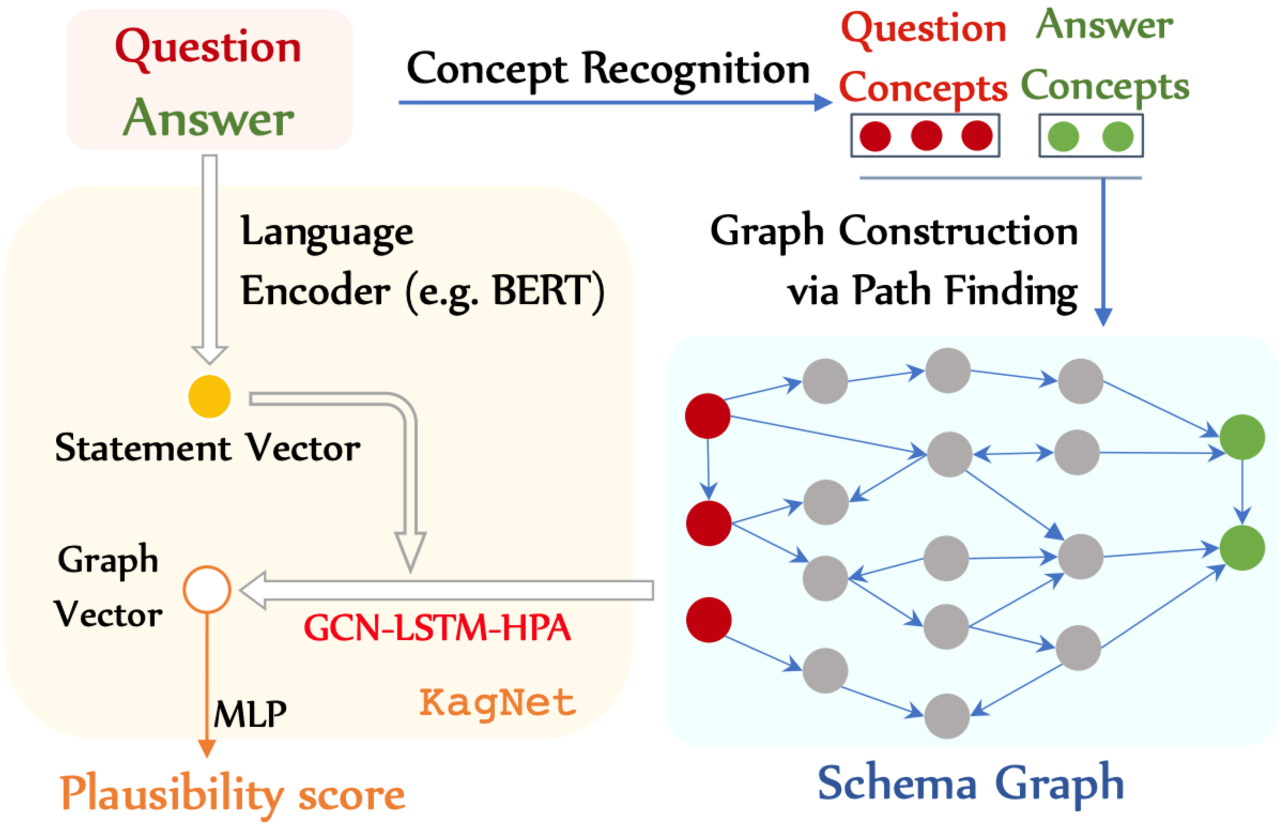
- Reporting bias => PTLM alone may not be sufficient

- 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

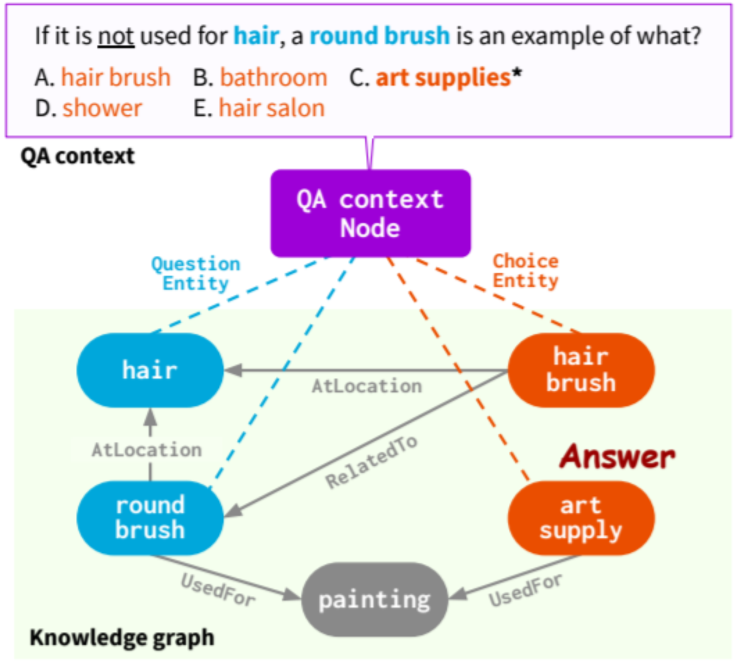
Q for Questions and A for Answers.

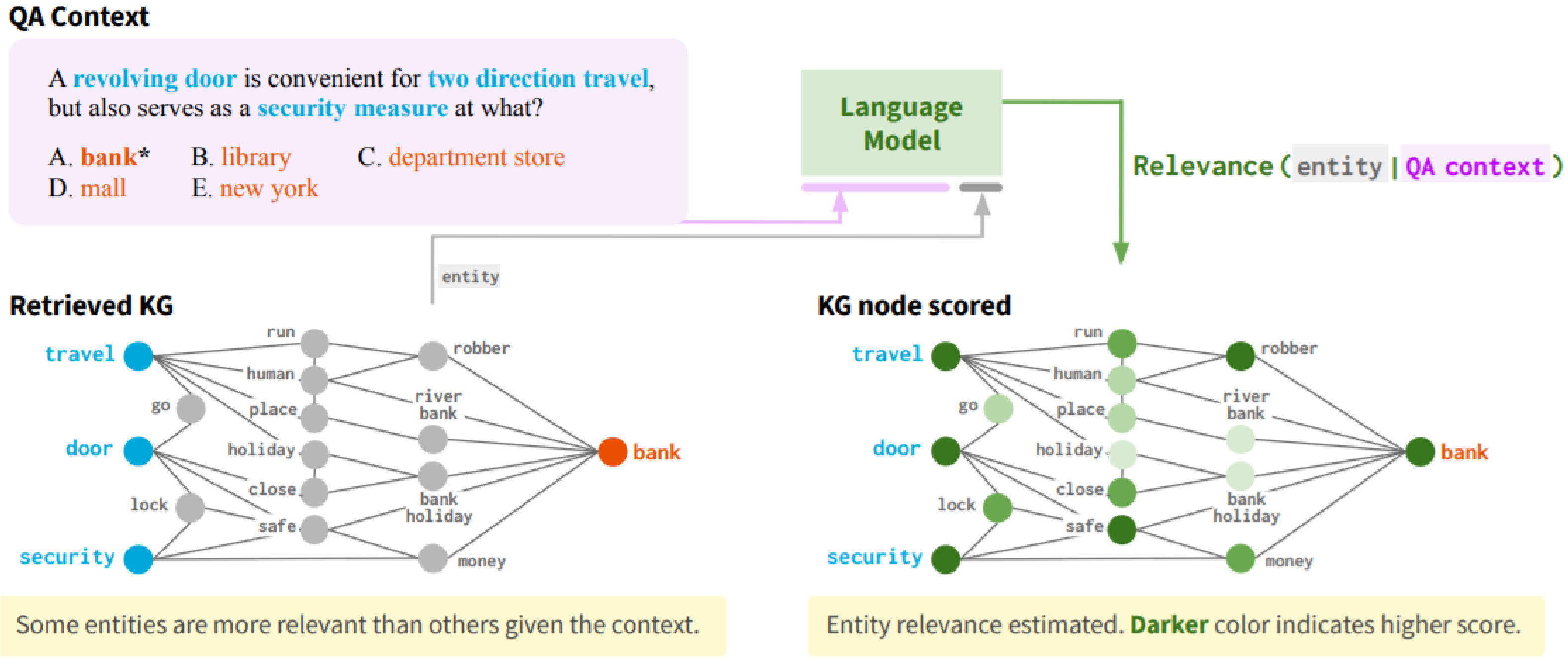


Lin, Bill Yuchen, et al. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. *EMNLP-IJCNLP*. 2019.

QA-GNN

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

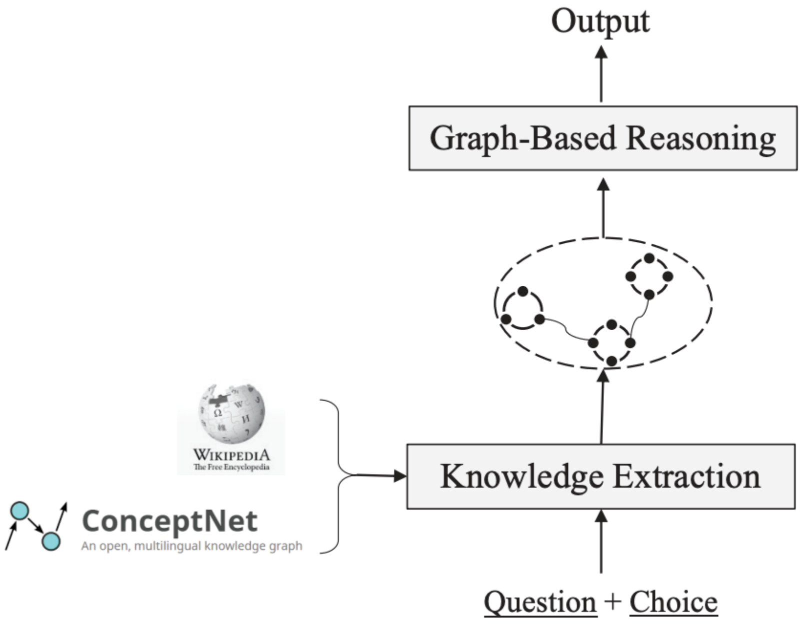


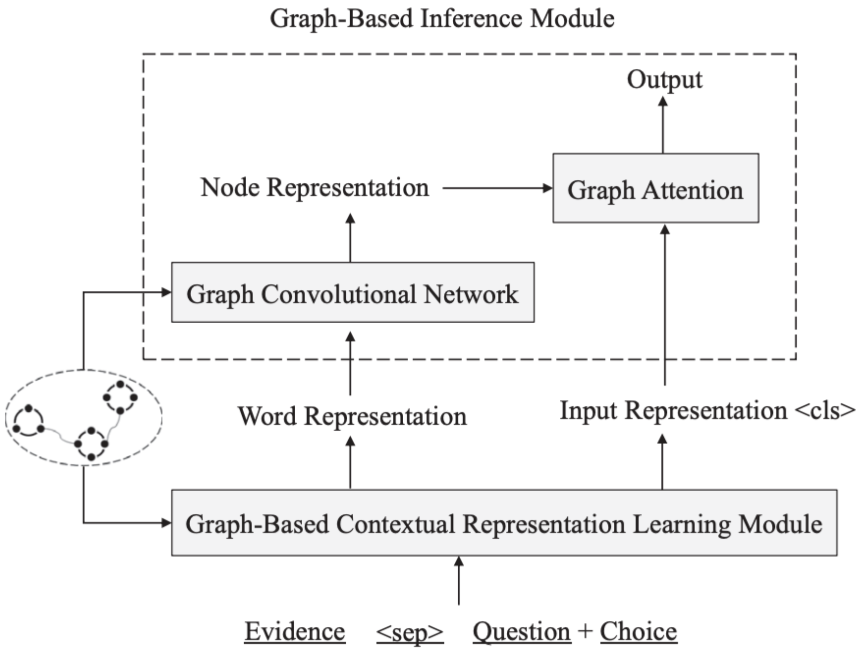


Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, Jure Leskovec . QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. NAACL, 2021.

ConceptNet+Wikipedia

- XLNet + Graph Reasoning
  - 1. Knowledge extraction (entity-based matchin) 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

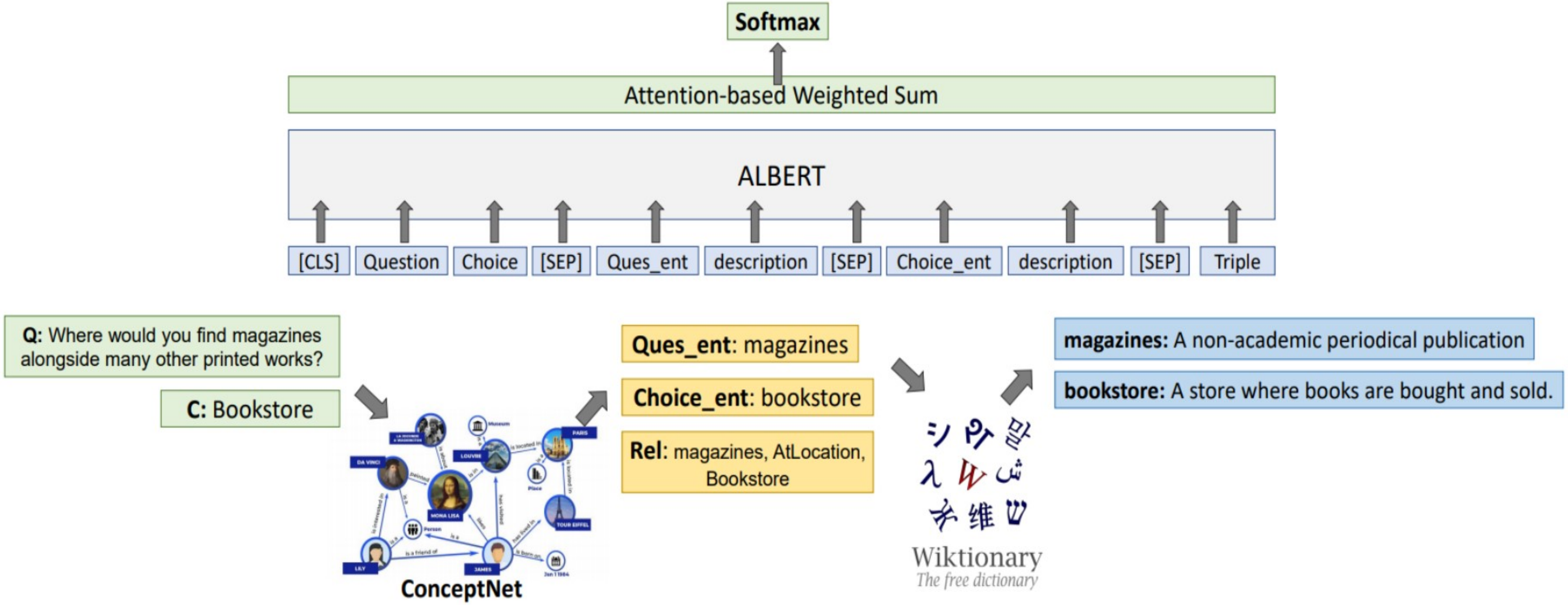




Lv, Shangwen, et al. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. AAAI 2020.

DEKCOR (Using Wiktionary Descriptions)

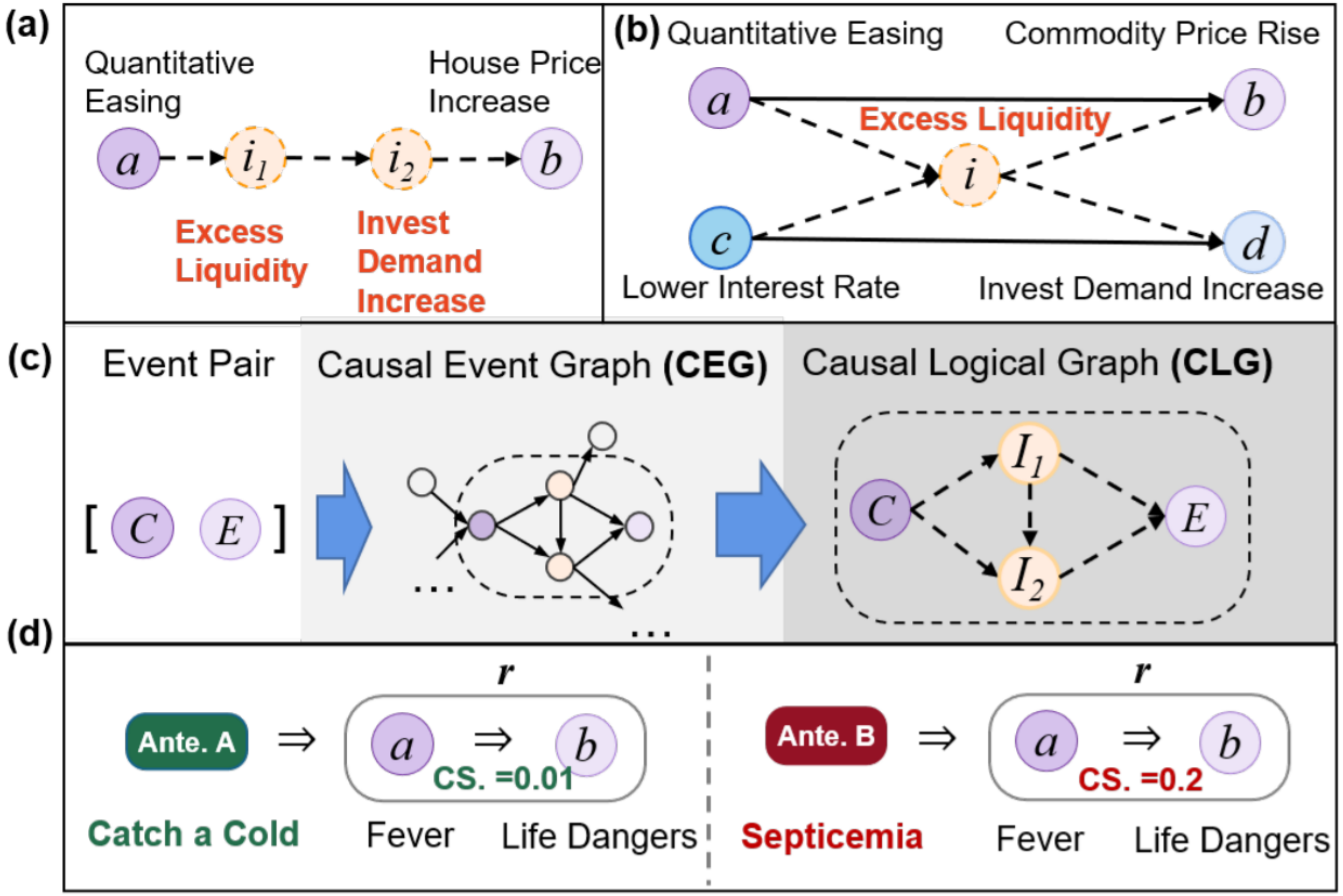
- 1. Retrieve ConceptNet subgraph.
- 2. Extract context (description of entities) from Wiktionary.
- 3. Reasoning (Attention)

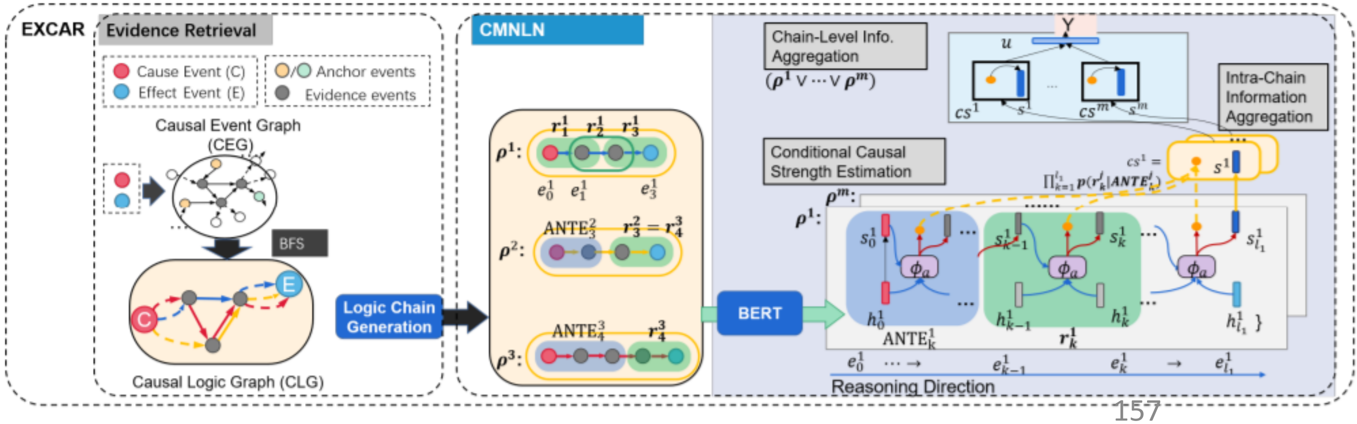


Xu, Yichong, et al. Fusing Context Into Knowledge Graph for Commonsense Reasoning. ACL 2021.

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





ExCAR: Event Graph Knowledge Enhanced Explainable Causal Reasoning Li Du, Xiao Ding∗ , Kai Xiong, Ting Liu, and Bing Qin

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

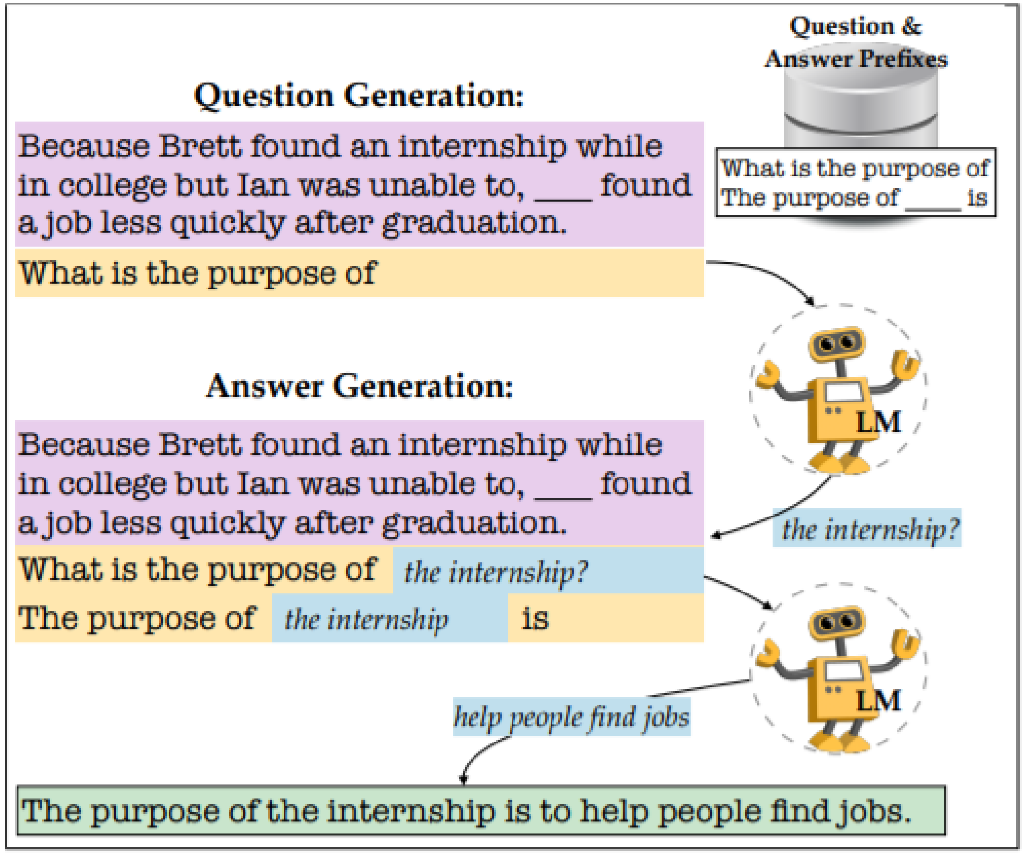
Induce Explicit Knowledge from Pre-Trained Language Model

- 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.

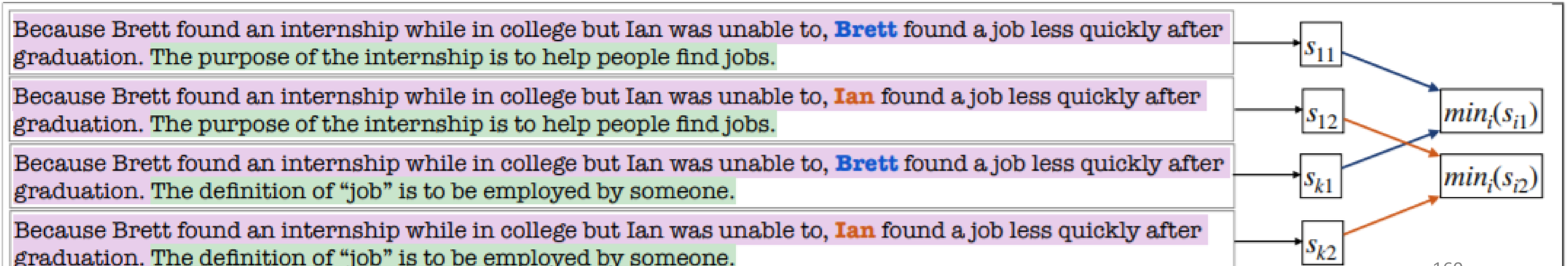
Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unsupervised commonsense question answering with self-talk. EMNLP, 2020.

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



Negative log likelihood

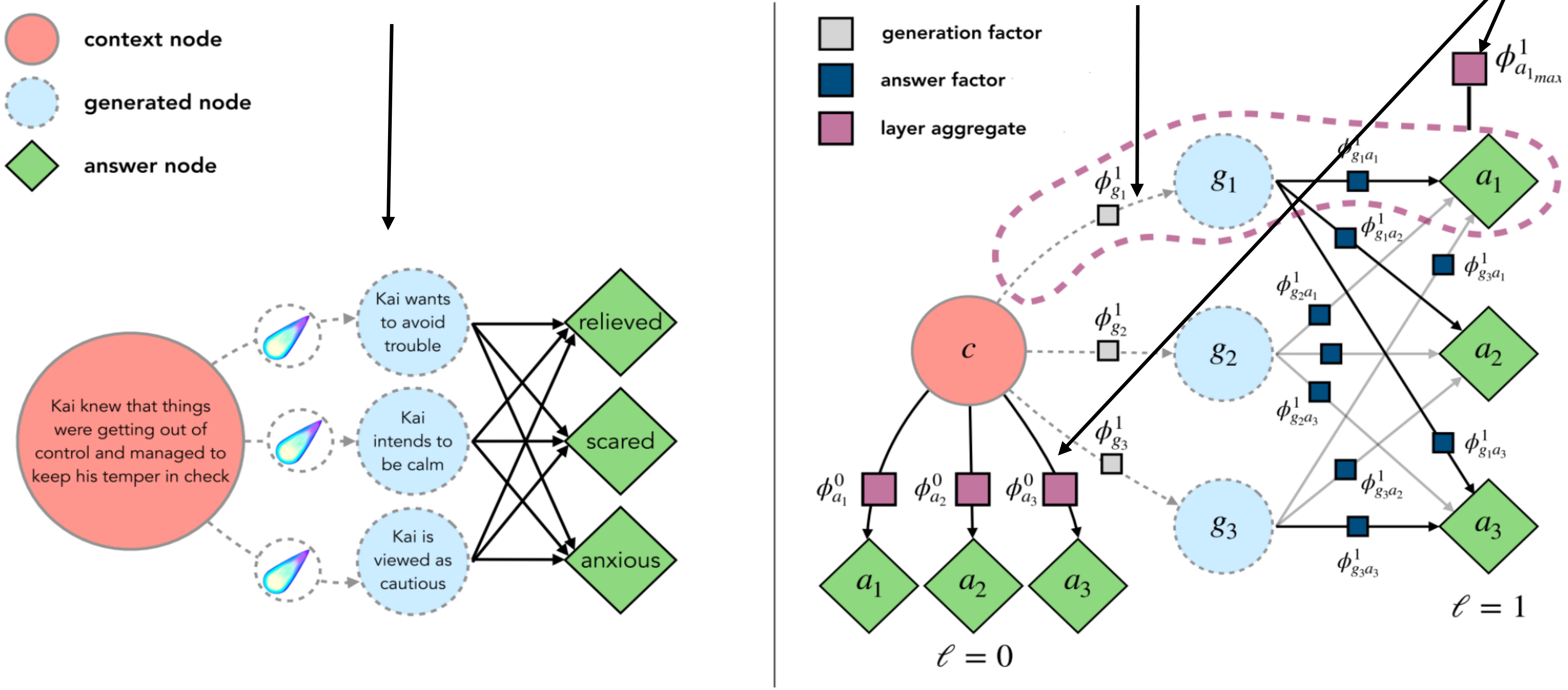
COMET-DynaGen (Bosselut et, al., 2019)

- 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)



Antoine Bosselut, Ronan Le Bras, and Yejin Choi. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. arXiv preprint arXiv:1911.03876, 2019.

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

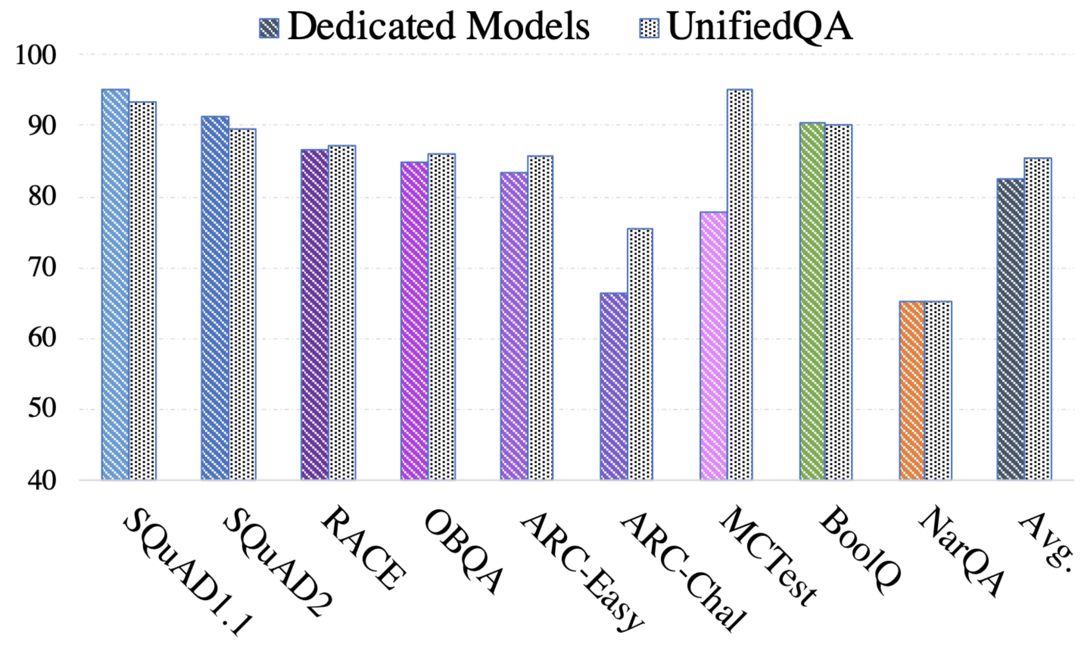
- Text-to-text unification:
  - Text in: [Question] + “\n” + ([Context], [Candidate Answers])
  - Text out: Answer
- Pre-trained on 8 QA datasets, SQuAD, NarrativeQA, RACE, ARC, etc.
  - Text-to-text PTLMs, BART and T5.
  - These pre-trained PTLM are then finetuned on each individual dataset for specific QAs.

EX	**Dataset**	SQuAD 1.1
	Input	At what speed did the turbine operate? \n (Nikola_Tesla) On his 50th birthday in 1906, Tesla demonstrated his 200 horsepower (150 kilowatts) 16,000 rpm bladeless turbine. ...
	Output	16,000 rpm
AB	**Dataset**	NarrativeQA
	Input	What does a drink from narcissus's spring cause the drinker to do? \n Mercury has awakened Echo, who weeps for Narcissus, and states that a drink from Narcissus's spring causes the drinkers to ``Grow dotingly enamored of themselves.'' ...
	Output	fall in love with themselves
MC	**Dataset**	ARC-challenge
	Input	What does photosynthesis produce that helps plants grow? \n (A) water (B) oxygen (C) protein (D) sugar
	Output	sugar
	Dataset	MCTest
	Input	Who was Billy? \n (A) The skinny kid (B) A teacher (C) A little kid (D) The big kid \n Billy was like a king on the school yard. A king without a queen. He was the biggest kid in our grade, so he made all the rules during recess. ...
	Output	The big kid
YN	**Dataset**	BoolQ
	Input	Was America the first country to have a president? \n (President) The first usage of the word president to denote the highest official in a government was during the Commonwealth of England ...
	Output	no

Khashabi, Daniel, et al. “UnifiedQA: Crossing Format Boundaries With a Single QA System.” *Findings of EMNLP*. 2020.

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



	CommonsenseQA	**WinoGrande**	**PIQA**	**SIQA**
BART-FT	62.5	62.4	77.4	74.0
UnifiedQA-BART-FT	64.0	63.6	77.9	73.2
T5-FT	78.1	84.9	88.9	81.4
UnifeidQA-T5-FT	**79.1**	**85.7**	**89.5**	**81.4**

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

	αNLI	**CosmosQA**	**HellaSWAG**	**PIQA**	**SIQA**	**WinoGrande**
multitask	78.4	81.1	81.3	80.7	74.8	72.1
finetune	79.2	82.6	**83.1**	**82.2**	75.2	78.2
sequential	**79.5**	**83.2**	83.0	**82.2**	**75.5**	**78.7**
none	77.8	81.9	82.2	80.2	73.8	77.0

Lourie, Nicholas, et al. UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark. AAAI, 2021.

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.

CSKG	**αNLI**	**CosmosQA**	**HellaSWAG**	**PIQA**	**SIQA**	**WinoGrande**
ATOMIC	**78.3**	81.8	**82.8**	79.9	**75.0**	**78.2**
ConceptNet	78.0	81.8	82.5	80.5	74.3	76.3
Both	78.0	81.8	82.7	**81.1**	74.8	76.6
Single Task	77.8	**81.9**	**82.8**	80.2	73.8	77.0

Summary of Results

	SWAG	SIQA	CosmosQA	PIQA	MCTACO	CommonsenseQA
Bert_{large}	86.6	64.5	66.8	66.7	42.72	56.7
XLNet_{large}	87.3	-	-	-	-	-
RoBERTa_{large}	89.9	78.7	81.9	79.4	54.8	72.1
ALBERT_{XXL}	**90.7**	-	85.4	-	-	**83.3**
T5_{11B}	-	81.4	90.3	88.9	-	78.1
UnifiedQA	-	81.5	-	89.5	-	79.1
UNICORN	-	**83.2**	**91.8**	**90.1**	-	79.3
COMET-DynaGen	-	52.6	-	-	-	-

- PTL
  -
  - Pre-training g
  - May not be sufficient for temporal CSQA yet

- UNICORN i
  - Multi-ta

- Self-Talk model can improve zero-shot learning
- BERT-large model has very low scores on several datasets:
  - Under-trained issue

Timeline of Approaches

CSKB Population

TransOMCS
Zhang, et al, 2020

DISCOS
Fang, et al, 2021

Benchmarking
Fang, et al, 2021

CSKB Completion

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

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

Neuro-Symbolic
KG Completion
Moghimifar, et al, 2021

Knowledge-enhanced

KagNet
Lin, et al 2019

HyKAS 2.0
Ma, et al 2019

XLNet+
Reason
Lv et al 2020

QA-GNN
Yasunaga, et al 2020

DESCKER
Xu, et al, 2021

2018 and before — 2019 — 2020 — 2021

Multi-task

UnifiedQA
Khashabi, et al, 2021

UNICORN
Lourie, et al, 2021

PTLM

GPT
Radford, et al, 2018

BERT
Delvin, et al, 2019

RoBERTa
Liu, et al, 2019

ALBERT
Lan, et al, 2020

BART
Lewis, et al, 2020

T5
Lourie, et al, 2020

DeBERTa
He, et al, 2020

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

$\alpha NLI/\alpha NLG$ 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

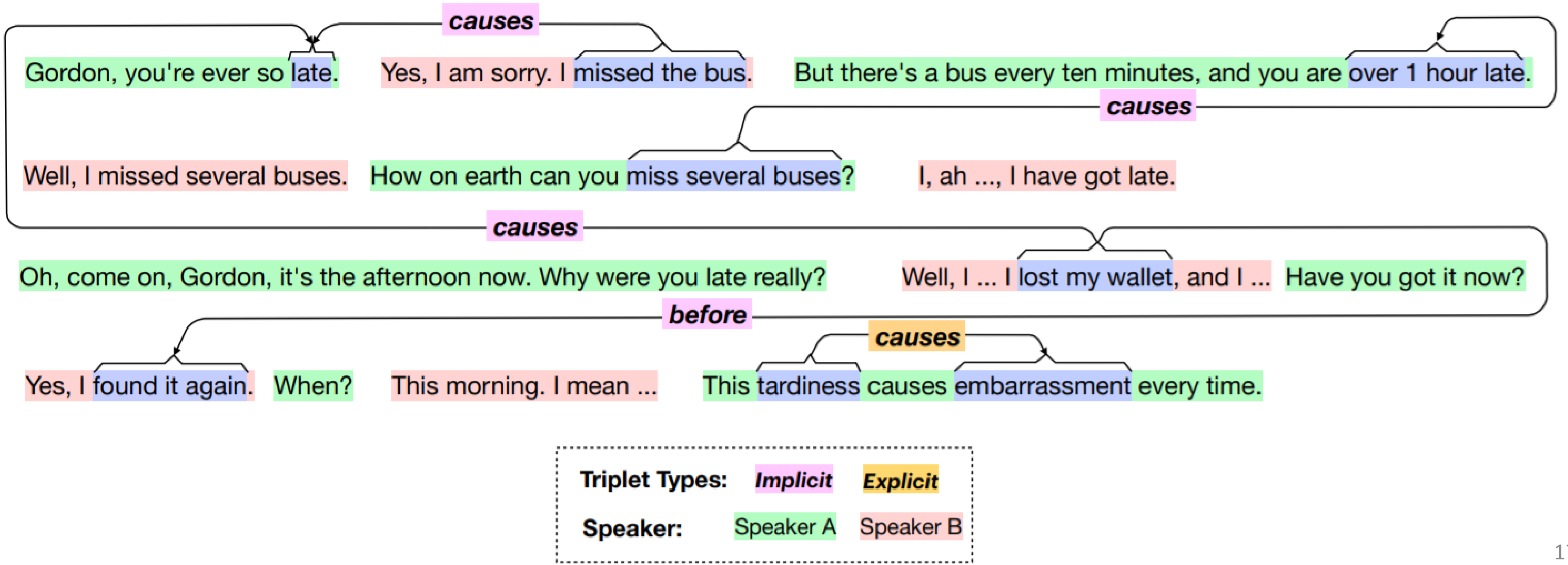
$$h^* = \arg\max_{h^i} P(H = h^i | O_1, O_2)$$

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

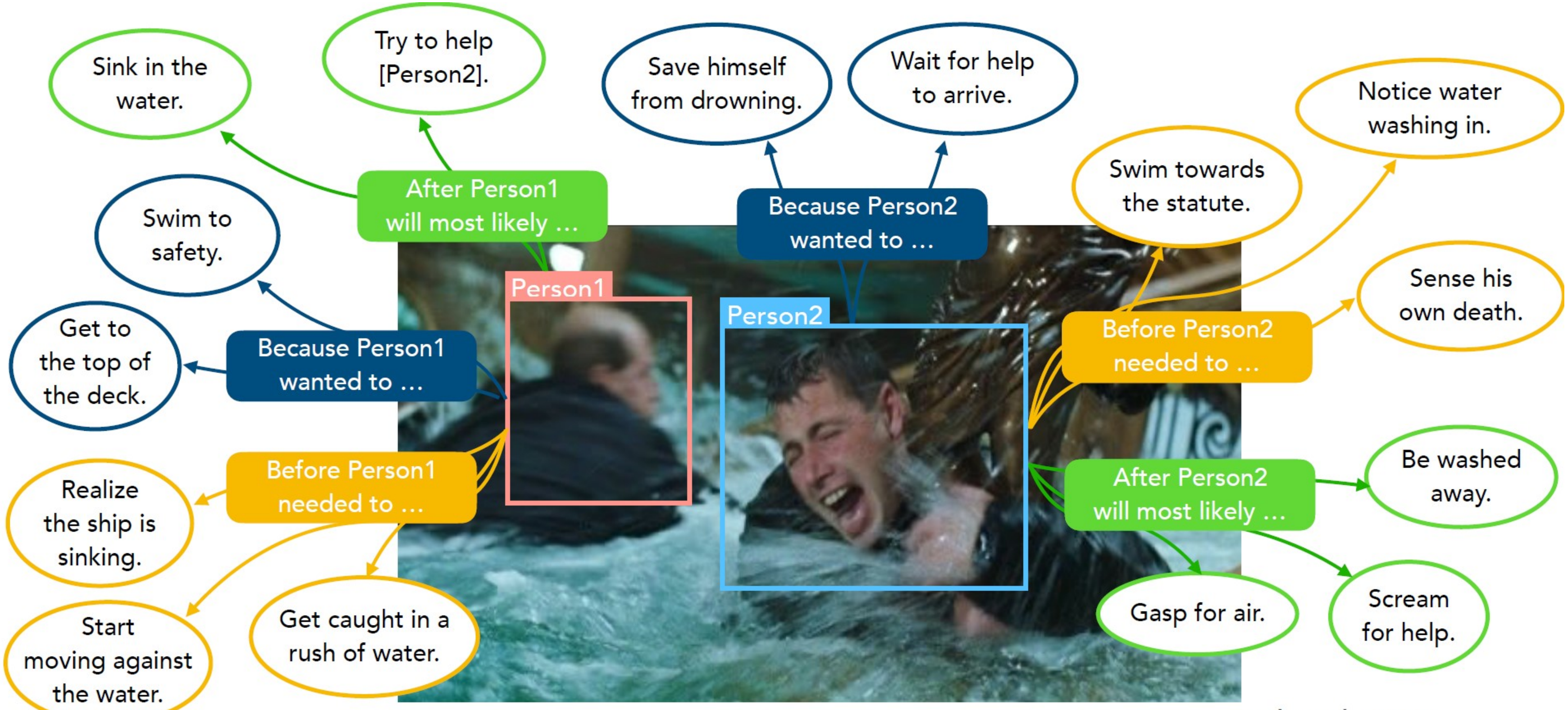
Abductive Commonsense Reasoning Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, Yejin Choi. ICLR, 2020.

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



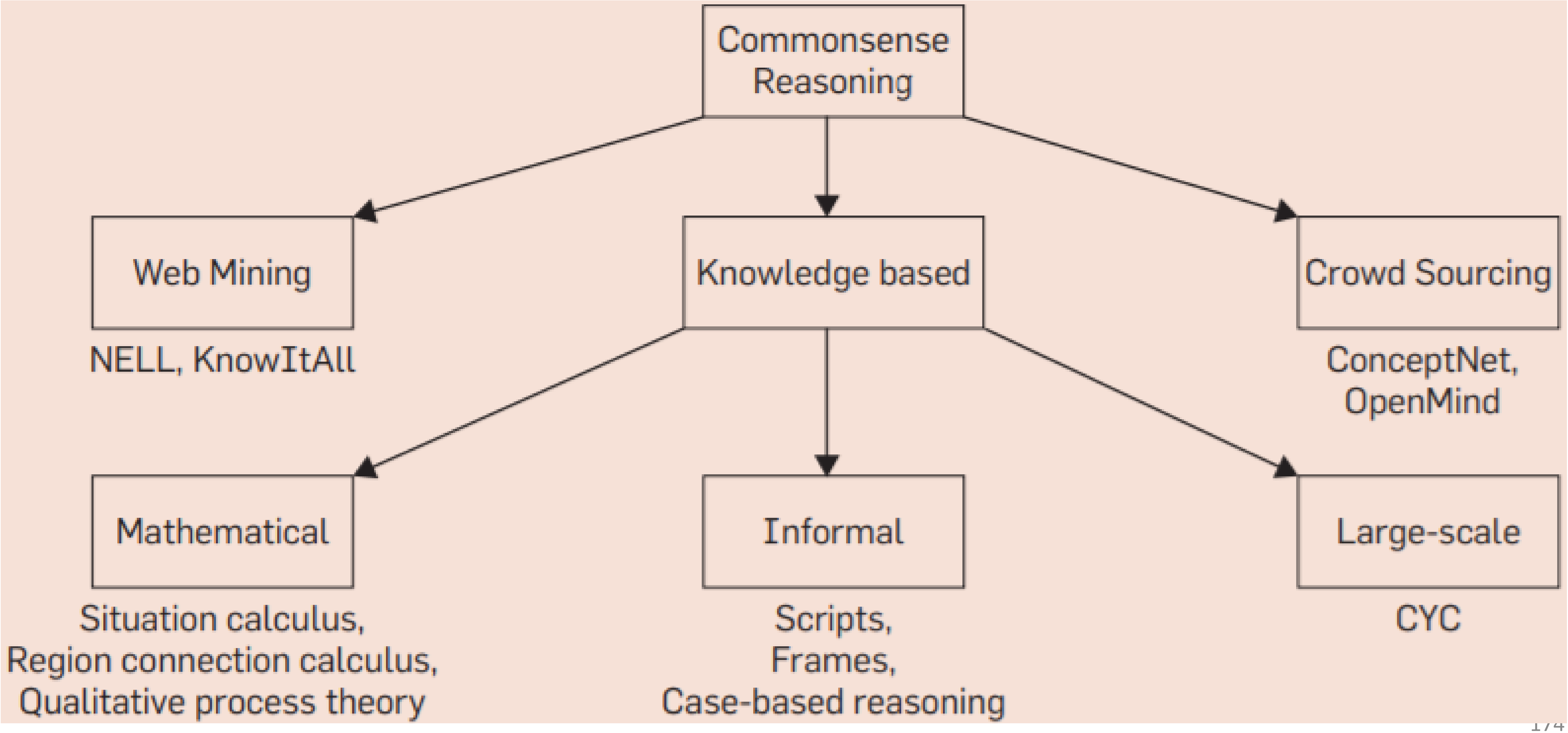
Park et al., 2020

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, Yejin Choi: VisualCOMET: Reasoning About the Dynamic Context of a Still Image. ECCV, 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 crowdsoucing 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!



Davis, Ernest, and Gary Marcus. "Commonsense reasoning and commonsense knowledge in artificial intelligence. " *Communications of the ACM* 58.9 (2015): 92-103.

Thank you for your attention! ☺