Commonsense Knowledge in the Open World

虞扬

Ref:

AAAI 2021 Tutorial on Commonsense Knowledge Acquisition and Representation

WSDM 2021 Information to Wisdom: Commonsense Knowledge Extraction and Compilation

Outline

- **What is CommonSense Knowledge (CSK)**
- Design Approach
- Extraction
- Consolidation
- Evaluation
- Forward

What is **C**ommon**S**ense **K**nowledge (CSK)

Definition

Definition 1 (Commonality)

Knowledge Shared by Nearly All Humans

Across culture and From early in life (~children)

Definition 2 (Type)

Knowledge about Concepts and Events

Concept: City, Footballer

Event: Football match, Birthday party

What is **C**ommon**S**ense **K**nowledge (CSK)

Definition

Definition 1 (Commonality)

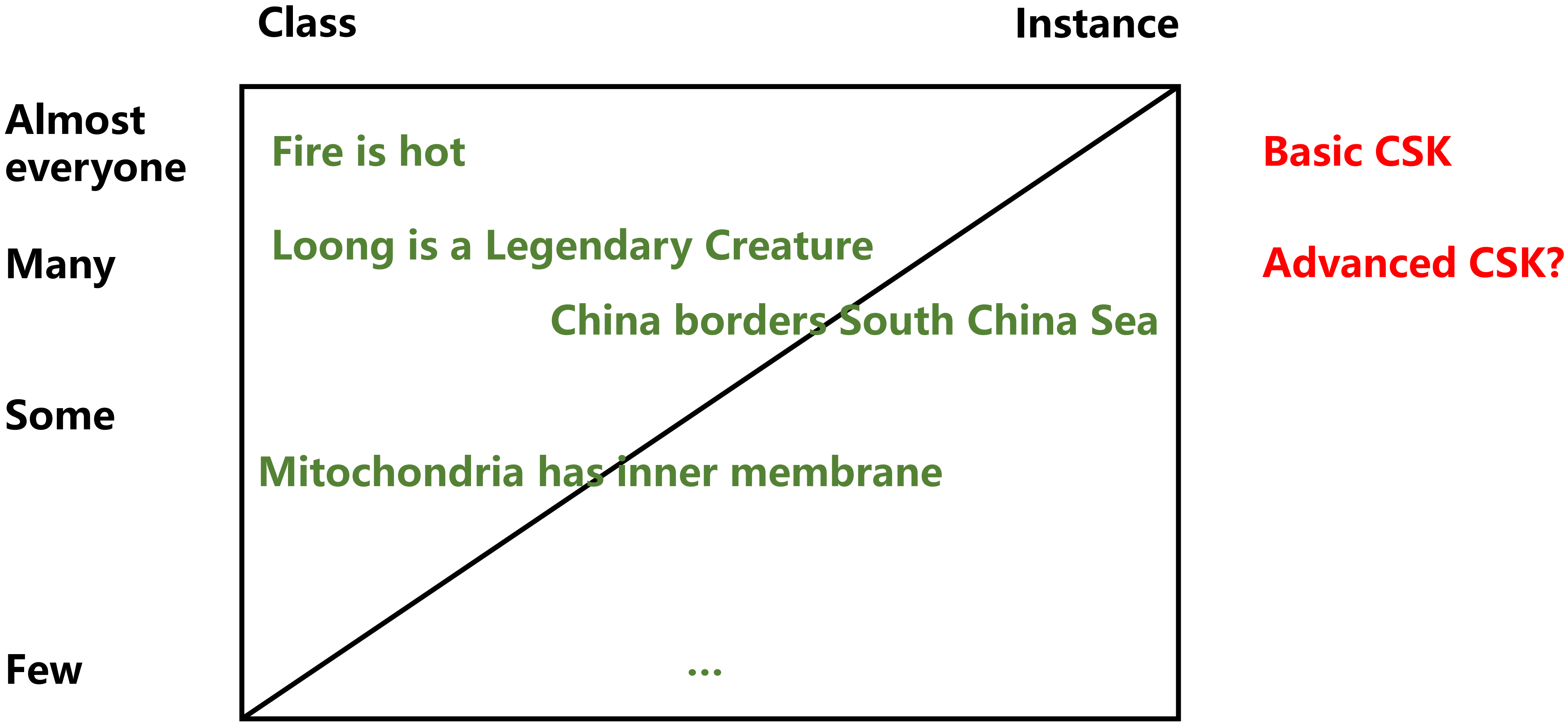
- Loong, isA, Legendary Creature – only known in Eastern Asian
- Lion is dangerous / cute – depend on whom you ask

Definition 2 (Type)

- Apple MacBook, Tesla Model X – instance
- China borders South China Sea – instance
- Mitochondria, hasPart, inner membrane – not common

What is CommonSense Knowledge (CSK)

Definition



What is **C**ommon**S**ense **K**nowledge (CSK)

Definition

- Taxonomical: **Loong , isA, Legendary Creature**
- Properties: **China, borders, South China Sea**
- Parts: **Elephants, hasPart, trunk**
- Measures: **Elephant, lifespan, ~60 years**
- Activities: **Go to zoo, subevent, Buy ticket**
- Causal: **Go to zoo, becauseOf, Seeing elephant**

…

What is **C**ommon**S**ense **K**nowledge (CSK)

Challenges

- Typically binary truth notion – **inf ∞**

- Across subjects

  Lion, has, manes **– percentage?**

- Corpus is subject to **reporting bias**
- Commonsense is not often written

  Grice’s maxim of quantity

Outline

- What is CommonSense Knowledge (CSK)
- **Design Approach**
- Extraction
- Consolidation
- Evaluation
- Forward

Design Approach

Top-Down: Axiom

Grandpa(Person, Person) |=

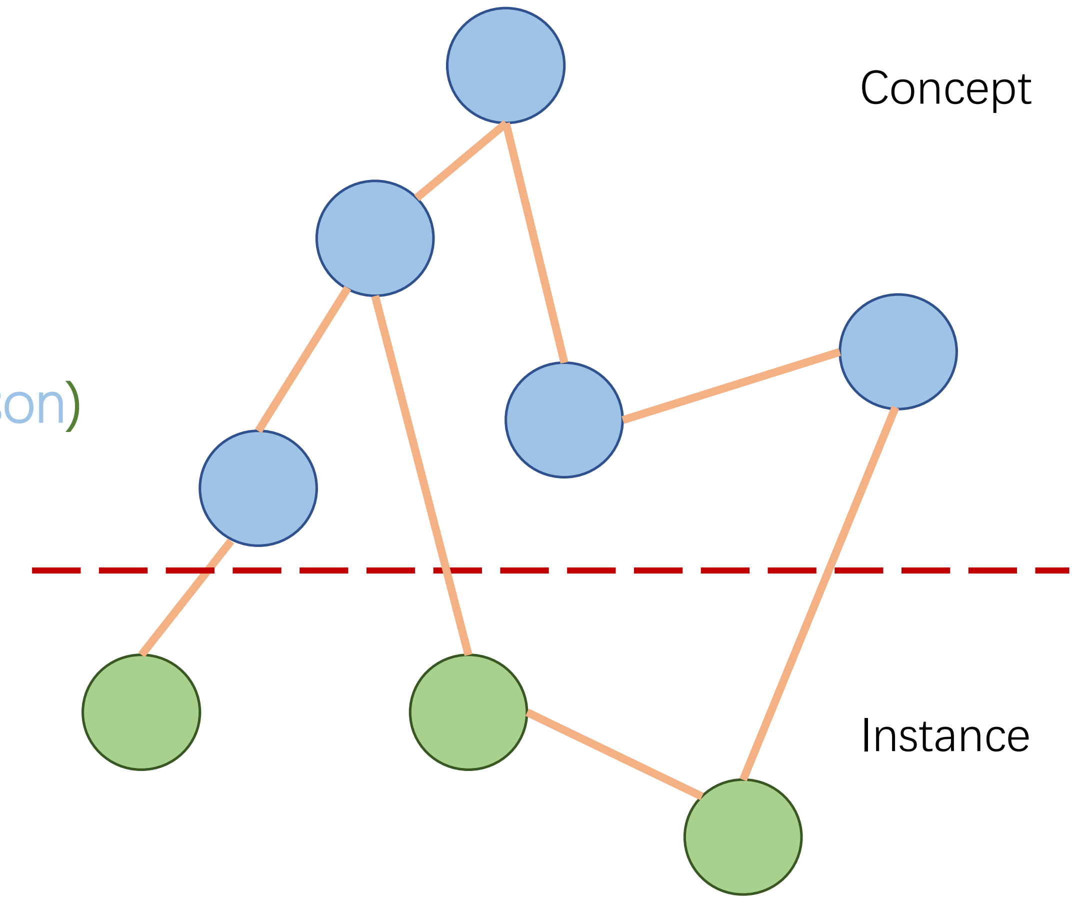
Father(Person, Person) ^ Father(Person, Person)

Down-Top: N-ary Tuple, Assertion

Fire, is, Hot

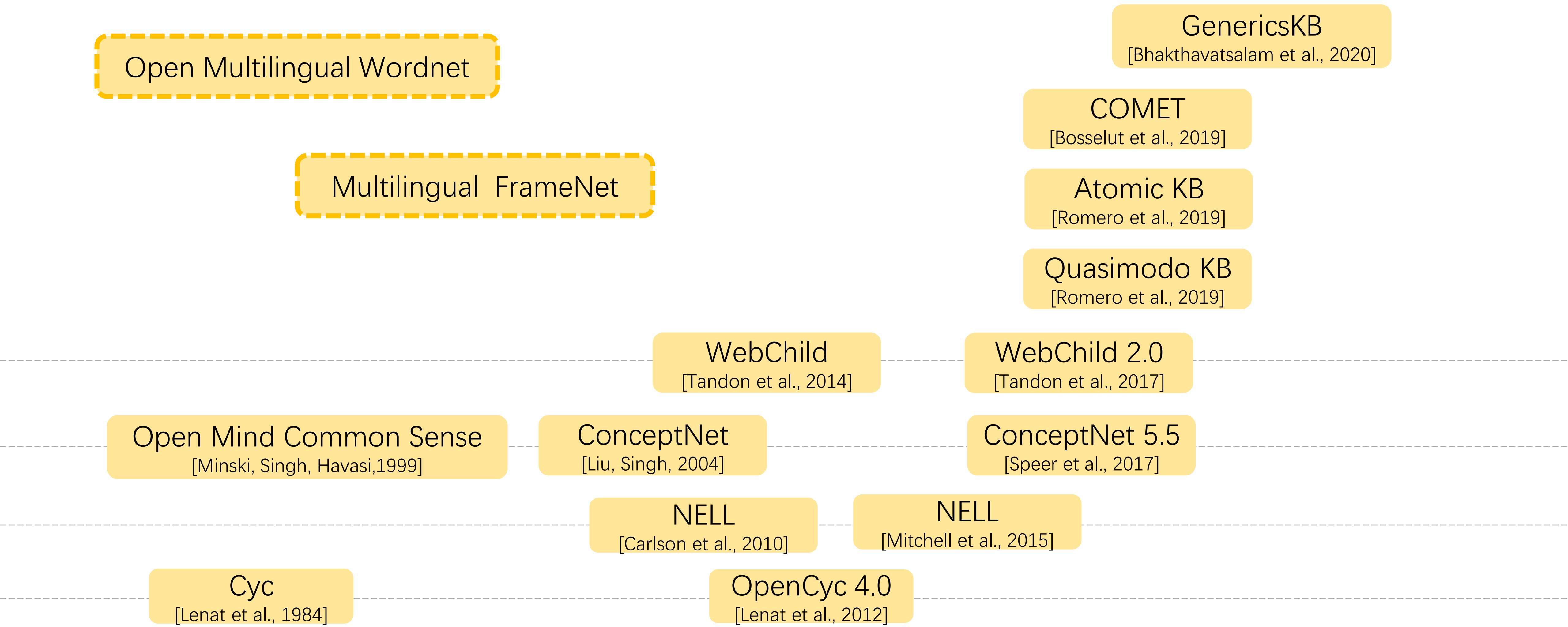
Lion, drink, Milk, childhood

Object Predicate Subject Others



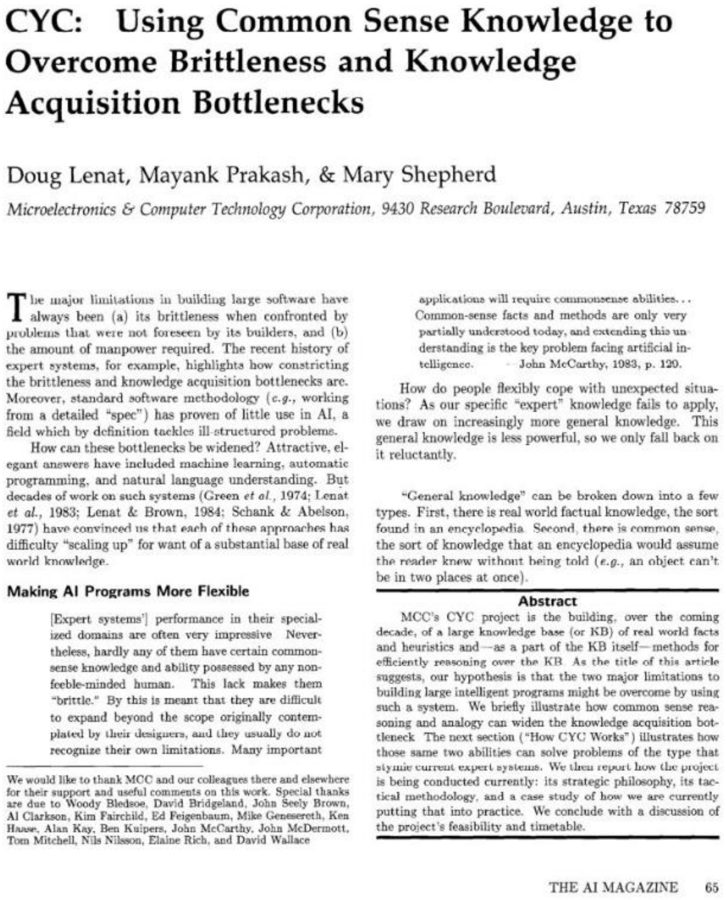
Design Approach

Down-Top – Example



Design Approach

Top-Down – Example



CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks

Doug Lenat, Mayank Prakash, & Mary Shepherd

Microelectronics & Computer Technology Corporation, 9430 Research Boulevard, Austin, Texas 78759

The major limitations in building large software have always been (a) its brittleness when confronted by problems that were not foreseen by its builders, and (b) the amount of manpower required. The recent history of expert systems, for example, highlights how constricting the brittleness and knowledge acquisition bottlenecks are. Moreover, standard software methodology (*e.g.*, working from a detailed "spec") has proven of little use in AI, a field which by definition tackles ill-structured problems.

How can these bottlenecks be widened? Attractive, elegant answers have included machine learning, automatic programming, and natural language understanding. But decades of work on such systems (Green *et al.*, 1974; Lenat *et al.*, 1983; Lenat & Brown, 1984; Schank & Abelson, 1977) have convinced us that each of these approaches has difficulty "scaling up" for want of a substantial base of real world knowledge.

Making AI Programs More Flexible

> [Expert systems'] performance in their specialized domains are often very impressive Nevertheless, hardly any of them have certain commonsense knowledge and ability possessed by any non-feeble-minded human. This lack makes them "brittle." By this is meant that they are difficult to expand beyond the scope originally contemplated by their designers, and they usually do not recognize their own limitations. Many important applications will require commonsense abilities... Common-sense facts and methods are only very partially understood today, and extending this understanding is the key problem facing artificial intelligence. —John McCarthy, 1983, p. 120.

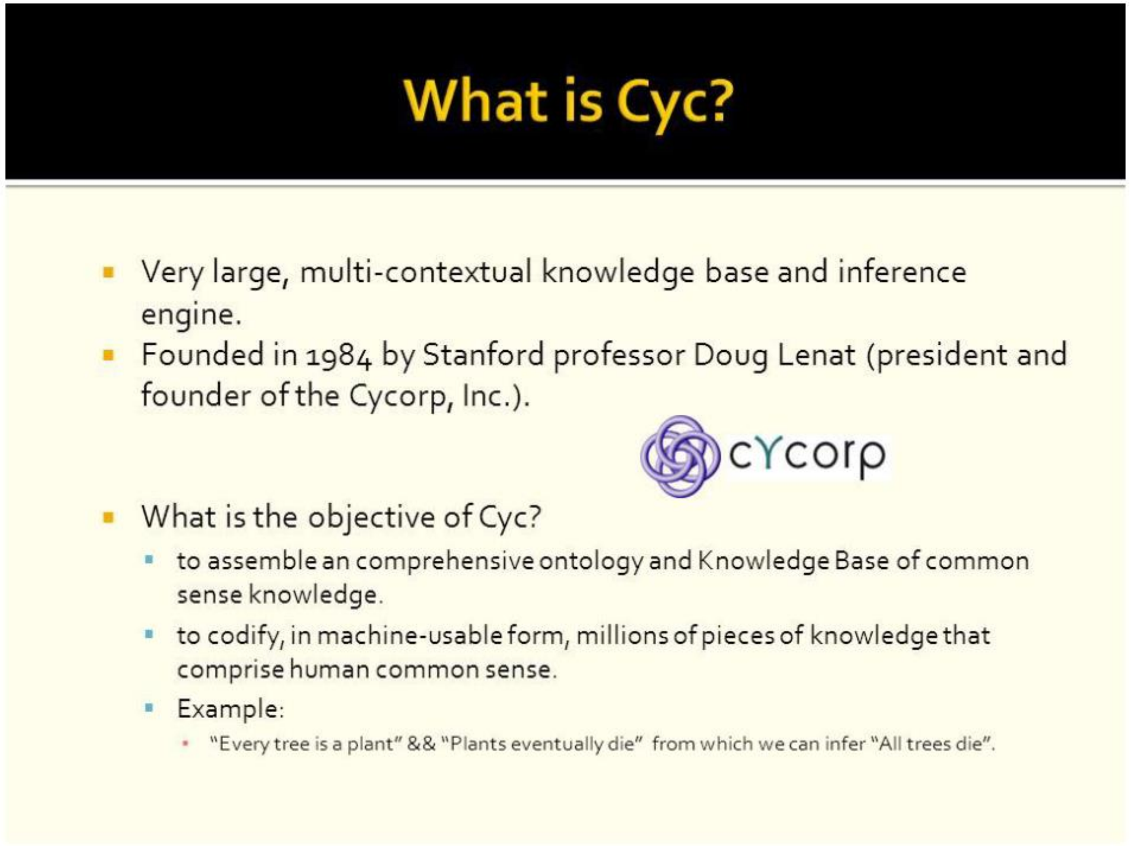
How do people flexibly cope with unexpected situations? As our specific "expert" knowledge fails to apply, we draw on increasingly more general knowledge. This general knowledge is less powerful, so we only fall back on it reluctantly.

"General knowledge" can be broken down into a few types. First, there is real world factual knowledge, the sort found in an encyclopedia Second, there is common sense, the sort of knowledge that an encyclopedia would assume the reader knew without being told (*e.g.*, an object can't be in two places at once).

Abstract

MCC's CYC project is the building, over the coming decade, of a large knowledge base (or KB) of real world facts and heuristics and—as a part of the KB itself—methods for efficiently reasoning over the KB As the title of this article suggests, our hypothesis is that the two major limitations to building large intelligent programs might be overcome by using such a system. We briefly illustrate how common sense reasoning and analogy can widen the knowledge acquisition bottleneck The next section ("How CYC Works") illustrates how those same two abilities can solve problems of the type that stymie current expert systems. We then report how the project is being conducted currently: its strategic philosophy, its tactical methodology, and a case study of how we are currently putting that into practice. We conclude with a discussion of the project's feasibility and timetable.

We would like to thank MCC and our colleagues there and elsewhere for their support and useful comments on this work. Special thanks are due to Woody Bledsoe, David Bridgeland, John Seely Brown, Al Clarkson, Kim Fairchild, Ed Feigenbaum, Mike Genesereth, Ken Haase, Alan Kay, Ben Kuipers, John McCarthy, John McDermott, Tom Mitchell, Nils Nilsson, Elaine Rich, and David Wallace



Outline

- What is CommonSense Knowledge (CSK)
- Design Approach
- **Extraction**
- Consolidation
- Evaluation
- Forward

Extraction

Source

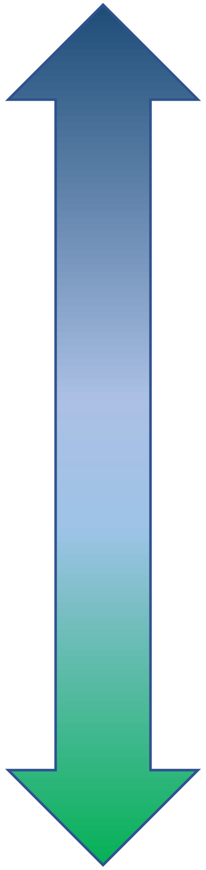
Wikipedia [Wikidata]

Topic Specific Knowledge

- Event: Wikihow [HowToKB, WWW 2017]
- Cultural: Movie scripts [Knowlywood, CIKM 2015]
- Science: Science textbooks [GenericsKB, Arxiv 2020]

Targeted Web Search [TupleKB, TACL 2017; Ascent, WWW, 2021]

Wild Web Pages [NELL, AAAI 2010]



Precision

Recall

Extraction

Textual Methods

- Manual Pattern [WebChild 2.0, ACL 2017]
- Co-occurrence [DoQ, ACL 2019]
- Open IE [TupleKB, Quasimodo, Ascent]

Extraction

Manual Pattern

Conventional method suffers from **Significant Noise:**

- **corpus**: Web-scale data, Web N-Gram dataset
- **seed:** ConceptNet (head, relation, tail)

Step 1) Generate Patterns per Relation:

e.g. “that apple is red”

pattern: <x> is <y>

Step 2) Score Pattern:

$$\phi(r,p) = \sum_{r' \in \mathcal{R}, r' \neq r} \frac{|S(r,p)|}{|S(r)|} - (1 - \mathrm{sim}(r,r')) \frac{|S(r',p)|}{|S(r')|}$$

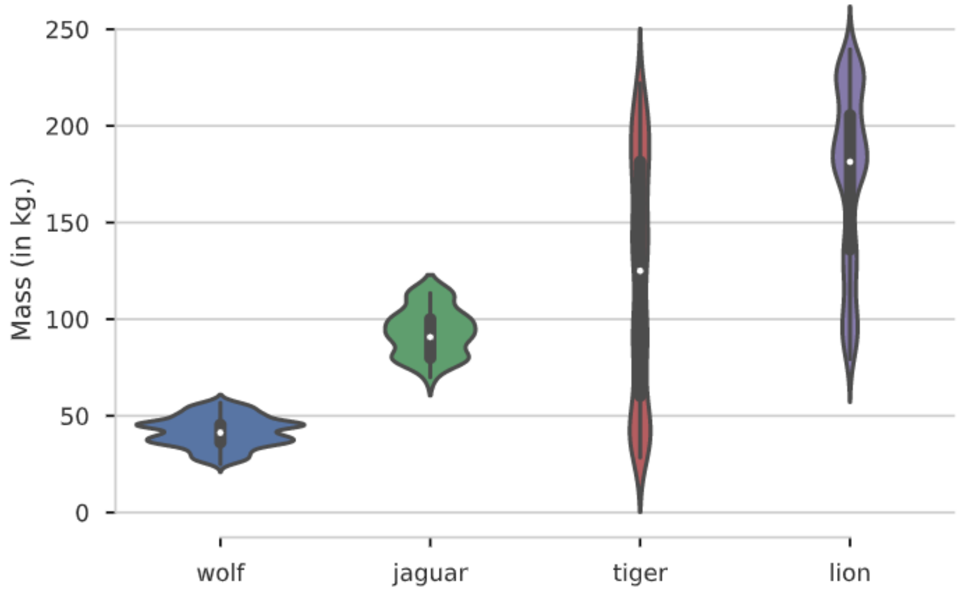
Step 3) Rank Assertion

WebChild 2.0: Fine-Grained Commonsense Knowledge Distillation. ACL. 2017

Extraction

Co-occurrence

Distribution over Quantities



(a) Mass distributions for multiple animals.

Noun Verb Noun

"These breeds can vary in weight from a

***0.46 kg** teacup poodle …"*

NP

460 gram

How Large Are Lions? Inducing Distributions over Quantitative Attributes. ACL. 2019

Digression

Co-occurrence

Noun *Verb* *Noun*

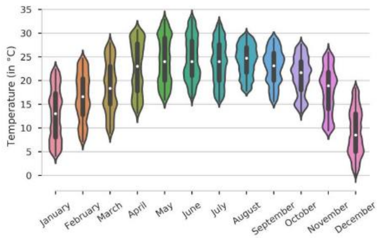
"These breeds can vary in weight from a

0.46 kg teacup poodle …"

NP

460 gram

- Measurement Detection
  rules: kg/kgs/kilogram/… -> normalize(kg->gram)
- Co-Occurring objects
  Pos Tagger (Nouns, Verbs, Adjectives …)
- Aggregating Measurements



How Large Are Lions? Inducing Distributions over Quantitative Attributes. ACL. 2019

Extraction

Open Information **E**xtraction (OpenIE)

OpenIE's goal is to read a sentence and extract *(arg1, relation, arg2)* with a relation *phrase* and *arguments* that are related by that relation phrase.

The U.S. president Barack Obama gave his speech on Tuesday to thousands of people.

(Barack Obama, is the president of, the U.S.)

(Barack Obama, gave, his speech)
(Barack Obama, gave his speech, on Tuesday)
(Barack Obama, gave his speech, to thousands of people)

Open Information Extraction Systems and Downstream Applications. IJCAI 2016.

Digression

Open Information **E**xtraction 5.1

Semantic Role Labeling (SRL)

Eli Whitney created the cotton gin in 1793

A_0 Verb A_1 Temporal

Conjunctive

Barack Obama visited India, Japan and South Korea.

(Barack Obama, visited, India)
(Barack Obama, visited, Japan),
(Barack Obama, visited, south Korea)

Numerical

Hong Kong' s labour force is 3.5 million.

(Hong Kong' s labour force, is, 3.5 million) (Hong Kong; has labour force of; 3.5 million)

Nominal

Japanese foreign minister Kishida

(Kishida, [is] foreign minister [of], Japan)

https://github.com/dair-iitd/OpenIE-standalone

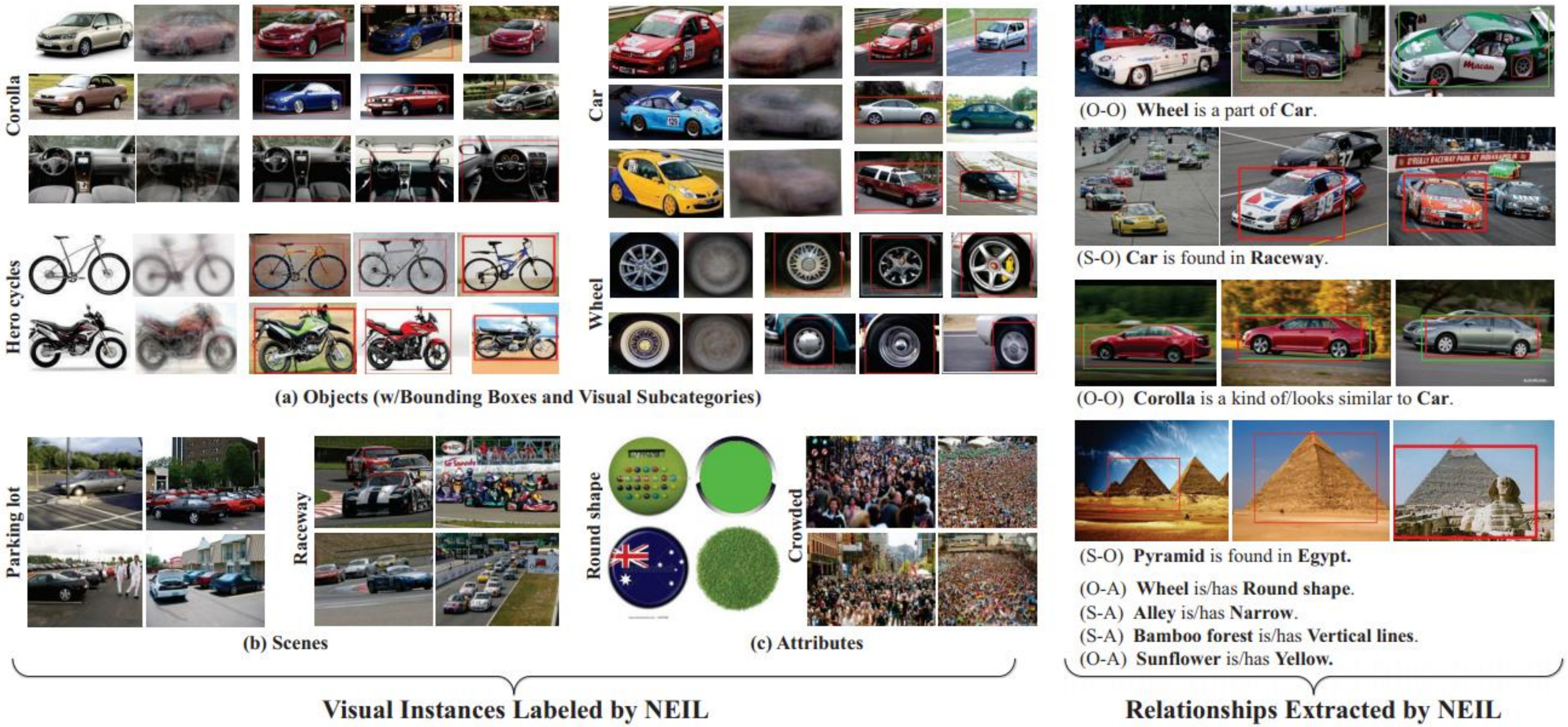
Open Information Extraction Systems and Downstream Applications. IJCAI 2016.

Extraction

Visual Method - NEIL

Knowledge:
1. Objects
2. Scenes
3. Attributes

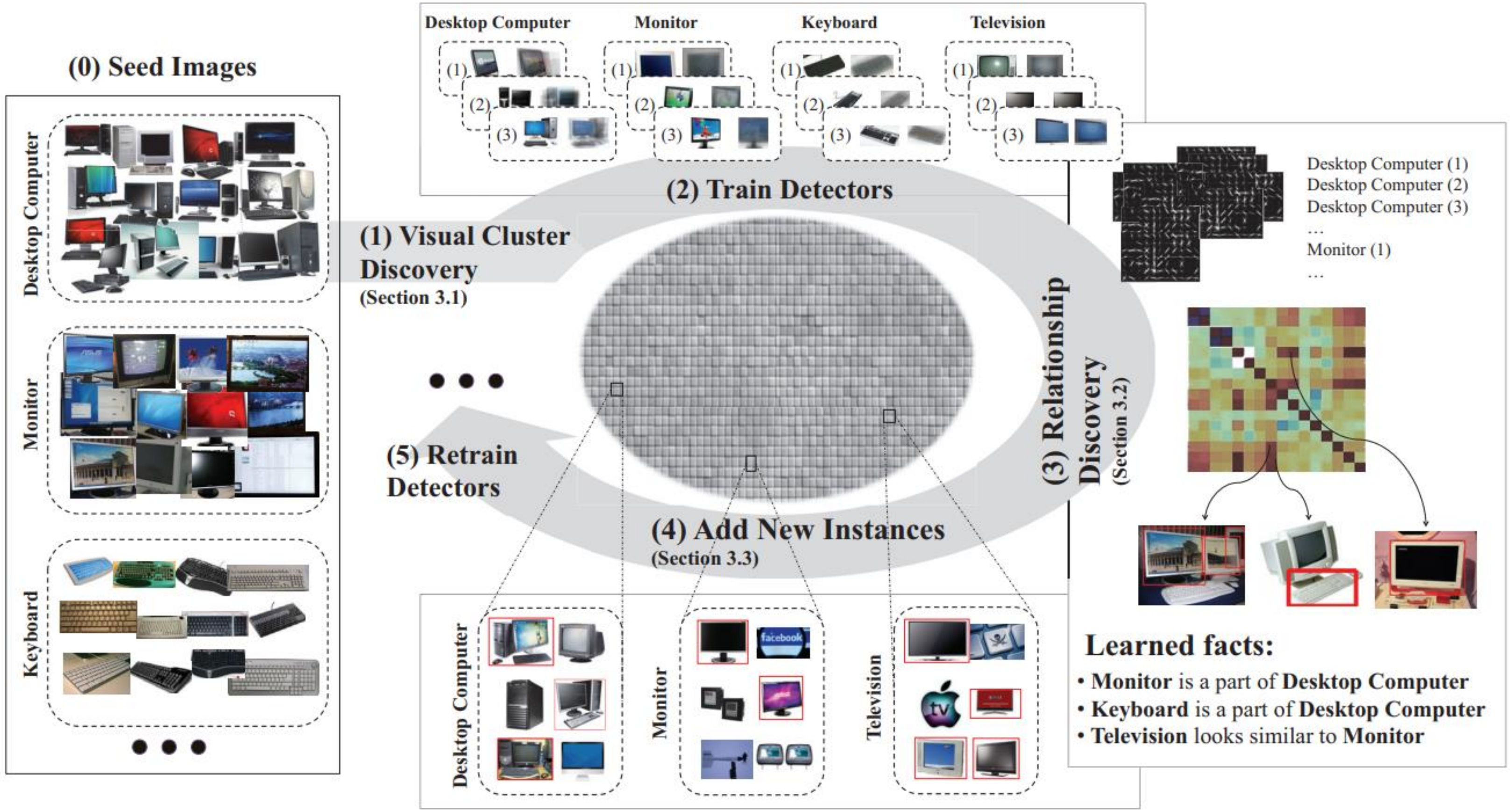
Relations
- A. Object-Object
- B. Object-Attribute
- C. Scene-Object
- D. Scene-Attribute



NEIL: Extracting Visual Knowledge from Web Data. ICCV. 2013

Extraction

Visual Method - NEIL



NEIL: Extracting Visual Knowledge from Web Data. ICCV. 2013

Outline

- What is CommonSense Knowledge (CSK)
- Design Approach
- Extraction
- **Consolidation**
- Evaluation
- Forward

Consolidation

Overview

Category	Source	Relations	Example 1	Example 2
Commonsense KGs	ConceptNet*	34	*food - capable of - go rotten*	*eating - is used for - nourishment*
	ATOMIC	9	*Person X bakes bread - xEffect - eat food*	*PersonX is eating dinner - xEffect - satisfies hunger*
	GLUCOSE	10	*Someone*$_A$ *makes Something*$_A$ *(that is food)*	*Causes/Enables Someone*$_A$ *eats Something*$_A$
	WebChild	4 (groups)	*restaurant food - quality#n#1 - expensive*	*eating - type of - consumption*
	Quasimodo	78,636	*pressure cooker - cook faster - food*	*herbivore - eat - plants*
	SenticNet	4	*cold_food - polarity - negative*	*eating breakfast - polarity - positive*
	HasPartKB	1	*dairy food - has part - vitamin*	*n/a*
Common KGs	Wikidata	6.7k	*food - has quality - mouthfeel*	*eating - subclass of - ingestion*
	YAGO4	116	*banana chip - rdf:type - food*	*eating - rdfs:label - feeding*
	DOLCE*	1	*n/a*	*n/a*
	SUMO*	1,614	*food - hyponym - food_product*	*process - subsumes - eating*
Lexical resources	WordNet	10	*food - hyponym - comfort food*	*eating - part-meronym - chewing*
	Roget	2	*dish - synonym - food*	*eating - synonym - feeding*
	FrameNet	8 (f2f)	*Cooking_creation - has frame element - Produced_food*	*eating - evoke - Ingestion*
	MetaNet	14 (f2f)	*Food - has role - food_consumer*	*consuming_resources - is - eating*
	VerbNet	36 (roles)	*feed.v.01 - Arg1-PPT - food*	*eating - hasPatient - comestible*
Visual sources	Visual Genome	42,374	*food - on - plate*	*boy - is eating - treat*
	Flickr30k	1	*a food buffet - corefers with - a food counter*	*a eating place - corefers with - their kitchen*
Corpora & LMs	GenericsKB	n/a	*Aardvarks search for food.*	*Animals receive nitrogen by eating plants.*
	GPT-2	n/a	*Food causes a person to be hungry and a person to eat.*	*Eating at home will not lead to weight gain.*

Consolidation

Hypothesis

piano is used for...

- performing music
- music
- accompanying an orchestra

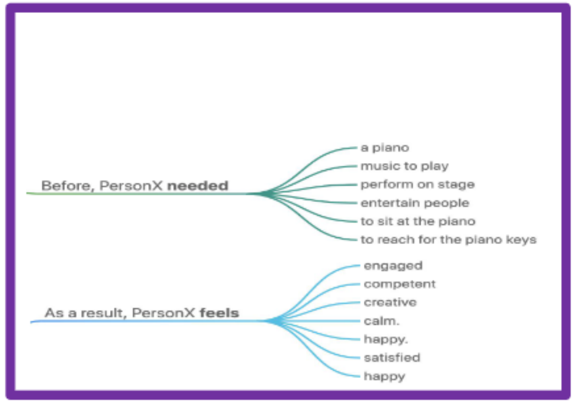
Things located at piano

- keys
- black keys
- hammers
- a keyboard

ConceptNet: pianos have keys, are used to perform music

- S: (n) **piano**, pianoforte, forte-piano (a keyboard instrument that is played by depressing keys that cause hammers to strike tuned strings and produce sounds)

WordNet: pianos are played by pressing keys



ATOMIC: to play piano, a person needs to sit at it, on stage and reach for the keys; feelings

On stage, a woman takes a seat at the piano. She

1. sits on a bench as her sister plays with the doll.
2. smiles with someone as the music plays.
3. is in the crowd, watching the dancers.
4. nervously sets her fingers on the keys.

FrameNet: performer entertains audience

Audience [Aud]	The Audience experiences the Performance.
Medium [Medium]	Medium is the physical entity or channel used by the Performer to transmit the Performance to the Audience.
Performance [Perance]	The Performers generates the Performance which the Audience perceives.
Performer [Perfer]	The Performer provides an experience for the Audience.

Visual Genome: person can play a piano while sitting, his hands are on the keyboard

- man plays piano
- keys ON piano
- woman watches man
- pillow ON couch
- light ON wall
- window IN room
- person playing piano
- guy ON bench
- hands ON keyboard

Consolidation

Challenges

- Knowledge granularity
- Imprecise descriptions
- Sparse overlap and mappings
- Modeling of relations

CSKG: The CommonSense Knowledge Graph. ArXiv. 2020

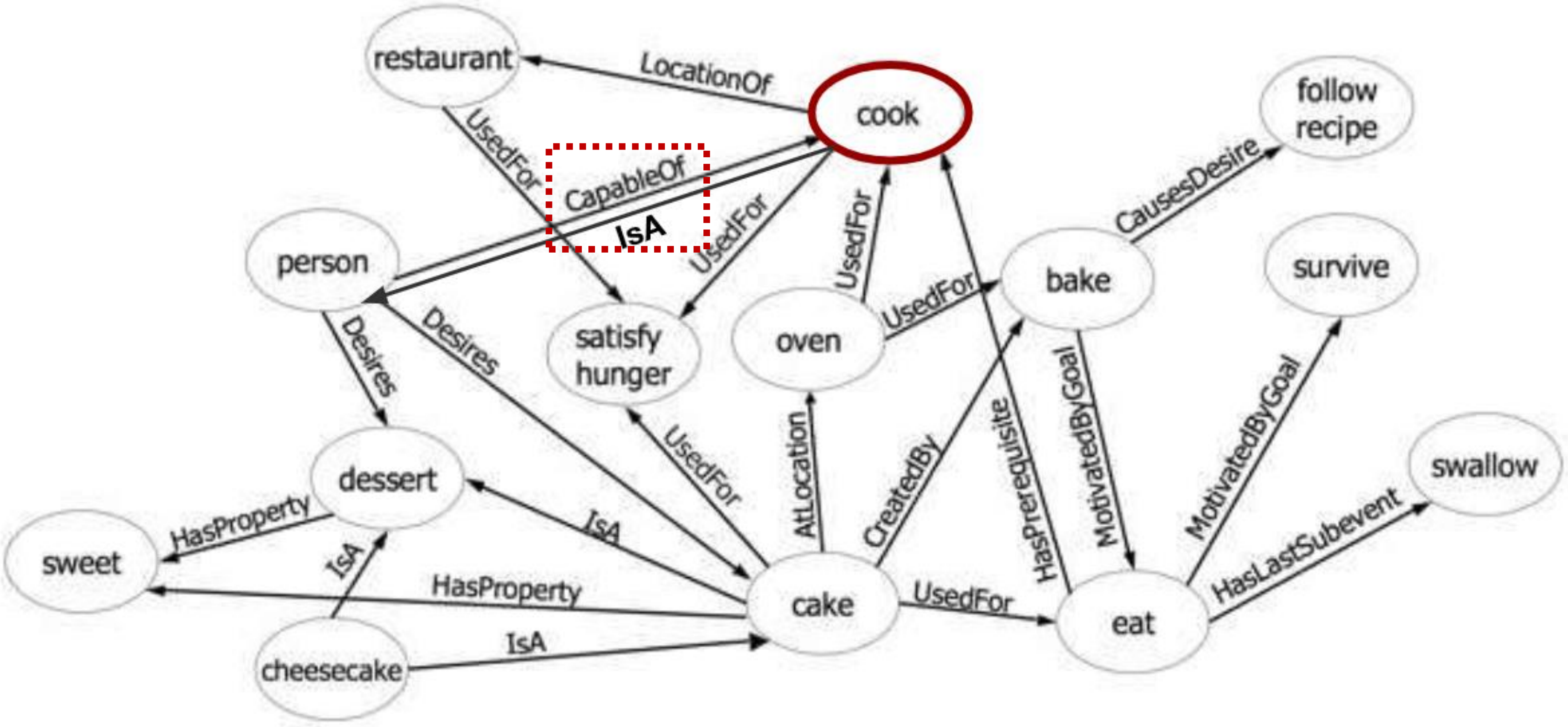
Consolidation

Knowledge granularity

	size	examples
Concept Net	36 relations, 8M nodes, 21M edges	/c/en/piano /c/en/piano/n /c/en/piano/n/wn /r/relatedTo
Web Child	4 relation groups, 2M nodes, 18M edges	hasTaste fasterThan
ATOMIC	9 relations, 300k nodes, 877k edges	wanted-to impressed
Wikidata	1.2k relations, 75M objects, 900M edges	wd:Q1234 wdt:P31
CEO	121 properties, 223 events	ceo:Damaging hasPostSituation
WordNet	10 relations, 155k words, 176k synsets	dog.n.01 hypernymy
Roget	2 relations, 72k words, 1.4M edges	truncate antonym
VerbNet	273 top classes 23 roles, 5.3k senses	perform-v performance-26.7-1
FrameNet	1.9k edges, 1.2k frames, 12k roles, 13k lexical units	Activity Change_of_leadership New_leader
Visual Genome	42k relations, 3.8M nodes, 2.3M edges, 2.8M attributes	fire hydrant white dog

Consolidation

Imprecise descriptions



Consolidation

Sparse overlap and mappings

	mappings
Concept Net	WordNet, DBpedia, OpenCyc, Wiktionary
Web Child	WordNet
ATOMIC	ConceptNet, Cyc
Wikidata	various
CEO	FrameNet, SUMO
WordNet	
Roget	
VerbNet	FrameNet, WordNet
FrameNet	
Visual Genome	WordNet

Consolidation

Modeling of relations

ConceptNet

/r/HasProperty

Web Child

-
ability#n#1
age#n#1
appearance#n#1
beauty#n#1
color#n#1
disposition#n#4
emotion#n#1
feeling#n#1
length#n#1
manner#n#1
motion#n#4
personality#n#1
physical_property#n#1
quality#n#1
sensitivity#n#2
shape#n#2
size#n#1
sound#n#1
state#n#2
strength#n#1
structure#n#2
sustainability#n#1
tactile_property#n#1
taste_property#n#1
temperature#n#1
trait#n#1
weight#n#1

Consolidation

Consolidate Nodes

P1. Embrace heterogeneity of nodes

objects, classes, words, actions, frames, states

P2. Leverage external links

many sources map to WordNet

P3. Generate high-quality probabilistic links

many facts not explicitly stated

Consolidation

Consolidate Relations

P1. Reuse edge types across resources -> 58 relations

/r/LocatedNear from ConceptNet applicable for attributes in Visual Genome

P2. Group relations into high-level dimensions

Causes, HasSubevent and precedes all express temporal knowledge

Dimension	ATOMIC	ConceptNet	WebChild	Other	Wikidata
lexical		FormOf DerivedFrom EtymologicallyDerivedFrom		lexical_unit (FN) lemma (WN)	label
temporal	xNeed xEffect oEffect xReact oReact	HasFirstSubevent HasLastSubevent HasSubevent HasPrerequisite Causes Entails	time emotion prev next	subframe (FN) precedes (FN) inchoative_of (FN) causative_of (FN)	has cause has effect
taxonomic		IsA InstanceOf MannerOf	hasHypernymy	perspective_on (FN) inheritance (FN) hypernym (WN)	subClassOf instanceOf description

...

Consolidation

Statistics

	AT	CN	FN	RG	WN	WD	VG	CSKG (concat)	CSKG
#nodes	304,909	1787373	15,652	71,804	91,294	71,243	11,264	2,414,813	2,160,968
#edges	732,723	3,423,004	54,109	1,403,955	111,276	101,771	2,587,623	6,349,731	6,001,531
#relations	9	34	23	2	3	15	3	59	58
mean degree	4.81	3.83	6.91	39.1	2.44	2.44	459.45	5.26	5.55

Outline

- What is CommonSense Knowledge (CSK)
- Design Approach
- Extraction
- Consolidation
- **Evaluation**
- Forward

Quality

No Ground Truth ☹

- Intrinsic Evaluation

  Is the Acquired Knowledge Good?

- Extrinsic Evaluation

  Is the Acquired Knowledge Useful?

Intrinsic Evaluation

Is the Acquired Knowledge Good?

Assess CSK-based systems to see how well they perform:

- Evaluate CSK systems: WebChild, TupleKB, DoQ, Quasimodo, Dice
- Important measures: Precision, Coverage
- Concept properties: Plausibility, Typicality, Remarkability, Saliency

Intrinsic Evaluation

Criteria

	Precision	Coverage	Plausibility	Typicality	Remarkability	Salience	Meaningfulness
WebChild	Y	Y					
TupleKB	Y	Y					
Quasimodo	Y	Y		Y		Y	Y
DoQ	Y	Y		Y			
Dice			Y	Y	Y	Y	

Precision: How correct is it? TP/(TP+FP)
Coverage (Recall): How much data does it cover? But there is no ground truth.

Plausibility(合理): Does the info make sense,
Typicality(典型): Is the info usually true? most lions eat meat
Remarkability(特性): Does the info stand out? hyenas eat carcasses
Salience(突出): Most distinguishing property? lions hunt in packs
Meaningfulness: Is it comprehensible?

Joint Reasoning for Multi-Faceted Commonsense Knowledge. AKBC. 2020

Intrinsic Evaluation

Gold Standard – Recall & Precision

Construct a data set of adjective-noun phrases labeled with appropriate attributes from WordNet 3.0

Assumption: examples given in glosses correspond to the respective word sense of the adjective.

- Extract all adjectives that are linked to at least one attribute synset by the attribute relation;
- Find examples of adjectives modifying nouns in attributive constructions; (TreeTagger)
- The resulting adjective-noun phrases are labeled with the attribute label linked to the given adjective sense;

Exploring supervised lDA models for assigning attributes to adjective-noun phrases. EMNLP. 2011

Intrinsic Evaluation

Gold Standard – Recall & Precision

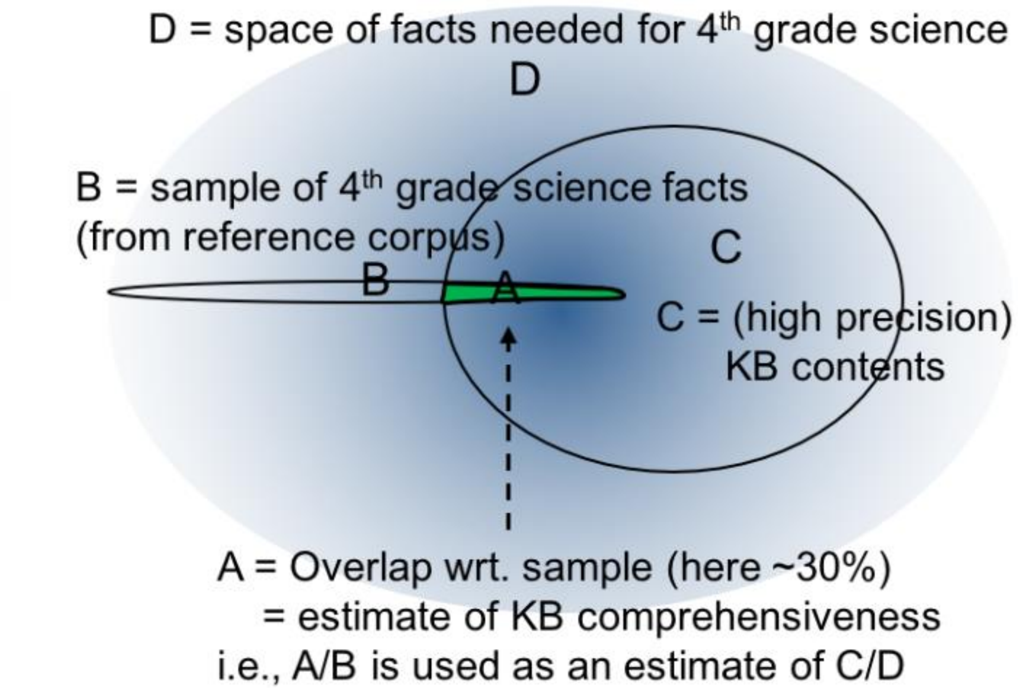
Method	Precision	Coverage
WordNet attributes	1.00	40
WordNet attributes expanded	0.61 ± 0.03	5,145
WordNet glosses	0.70 ± 0.06	3,698
Controlled LDA MFS	0.30 ± 0.06	2,775
Google Sets MFS	0.27 ± 0.04	426
WebChild	0.90 ± 0.03	7,783

WebChild: harvesting and organizing commonsense knowledge from the web. WSDM. 2014

Intrinsic Evaluation

Comprehensiveness – Coverage

Definition: recall at high (>80%) precision of domain-relevant facts



KB	Precision	Coverage of Tuple-Expressible Science Knowledge (Recall on science KB)
WebChild	89%	3.4%
NELL	85%	0.1%
ConceptNet	40%	8.4%
ReVerb-15M	55%	11.5%
Our KB	81%	23.2%

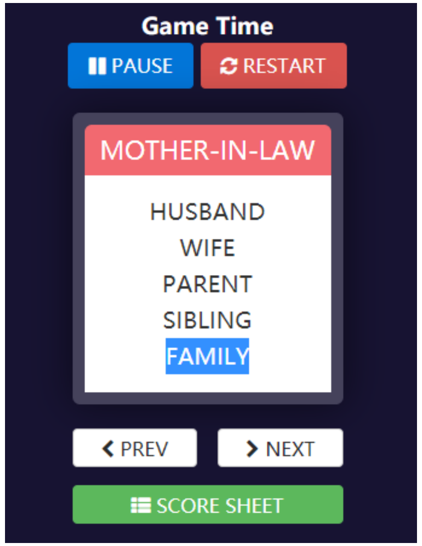
Domain-targeted, high precision knowledge extraction. TACL, 2017

Intrinsic Evaluation

Typicality & Remarkability & Recall

Quasimodo: CSK related to QA

Idea: Questions convey salient knowledge



Taboo Game

Crowdsourcing Task

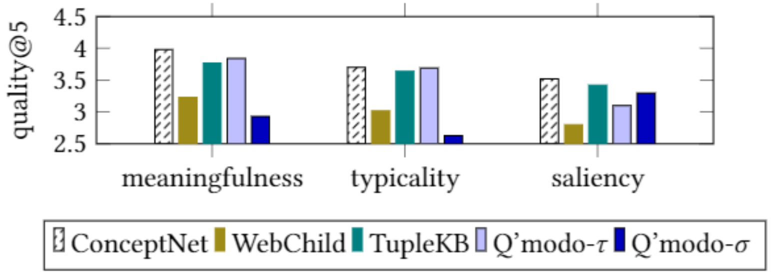


Figure 4: Quality for horizontal sampling.

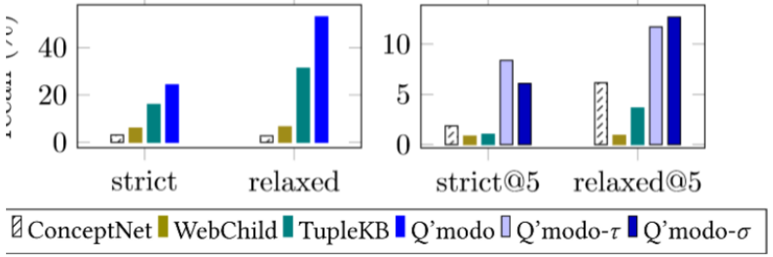


Figure 5: Recall evaluation.

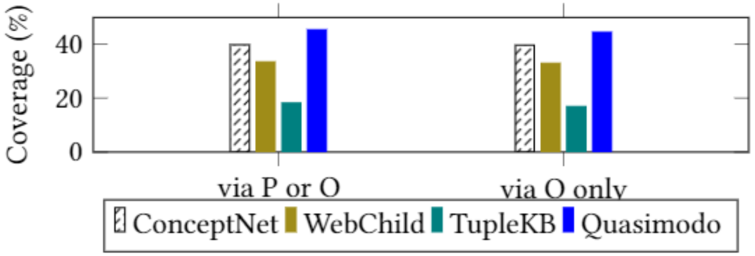


Figure 6: Coverage for word guessing game.

Commonsense Properties from Query Logs and Question Answering Forums. ArXiv. 2019

Intrinsic Evaluation

DIverse **C**ommonsense **K**nowledge (DICE)

Let $\mathcal{S}$ denote the set of subjects and $\mathcal{P}$ the properties

Concept-dimension dependencies: $\forall(s,p) \in \mathcal{S} \times \mathcal{P}$

$$\text{Typical}(s,p) \Rightarrow \text{Plausible}(s,p) \quad (1)$$

$$\text{Salient}(s,p) \Rightarrow \text{Plausible}(s,p) \quad (2)$$

$$\text{Typical}(s,p) \wedge \text{Remarkable}(s,p) \Rightarrow \text{Salient}(s,p) \quad (3)$$

Parent-child dependencies: $\forall(s_1,p) \in \mathcal{S} \times \mathcal{P}, \forall s_2 \in \text{children}(s_1)$

$$\text{Plausible}(s_1,p) \Rightarrow \text{Plausible}(s_2,p) \quad (4)$$

$$\text{Typical}(s_1,p) \Rightarrow \text{Typical}(s_2,p) \quad (5)$$

$$\text{Typical}(s_2,p) \Rightarrow \text{Plausible}(s_1,p) \quad (6)$$

$$\text{Remarkable}(s_1,p) \Rightarrow \neg\text{Remarkable}(s_2,p) \quad (7)$$

$$\text{Typical}(s_1,p) \Rightarrow \neg\text{Remarkable}(s_2,p) \quad (8)$$

$$\neg\text{Plausible}(s_1,p) \wedge \text{Plausible}(s_2,p) \Rightarrow \text{Remarkable}(s_2,p) \quad (9)$$

$$(\forall s_2 \in \text{children}(s_1)\ \ \text{Typical}(s_2,p)) \Rightarrow \text{Typical}(s_1,p) \quad (10)$$

Sibling dependencies: $\forall(s_1,p) \in \mathcal{S} \times \mathcal{P}, \forall s_2 \in \text{siblings}(s_1)$

$$\text{Remarkable}(s_1,p) \Rightarrow \neg\text{Remarkable}(s_2,p) \quad (11)$$

$$\text{Typical}(s_1,p) \Rightarrow \neg\text{Remarkable}(s_2,p) \quad (12)$$

$$\neg\text{Plausible}(s_1,p) \wedge \text{Plausible}(s_2,p) \Rightarrow \text{Remarkable}(s_2,p) \quad (13)$$

Joint Reasoning for Multi-Faceted Commonsense Knowledge. AKBC. 2020

Extrinsic Evaluation

Is the Acquired Knowledge Useful ?

- **Utility of the knowledge**
- Comprehension **benchmarks & applications** with extrinsic use cases to test CSK systems

Linguistic reasoning commonsense (text)

WinoGrande: Large scale dataset as commonsense reasoning benchmark in language [arXiv 2019]
DoQA: Domain-specific conversational QA for FAQs [ACL 2020]
Arc: Ai2 Reasoning Challenge [arXiv 2018]

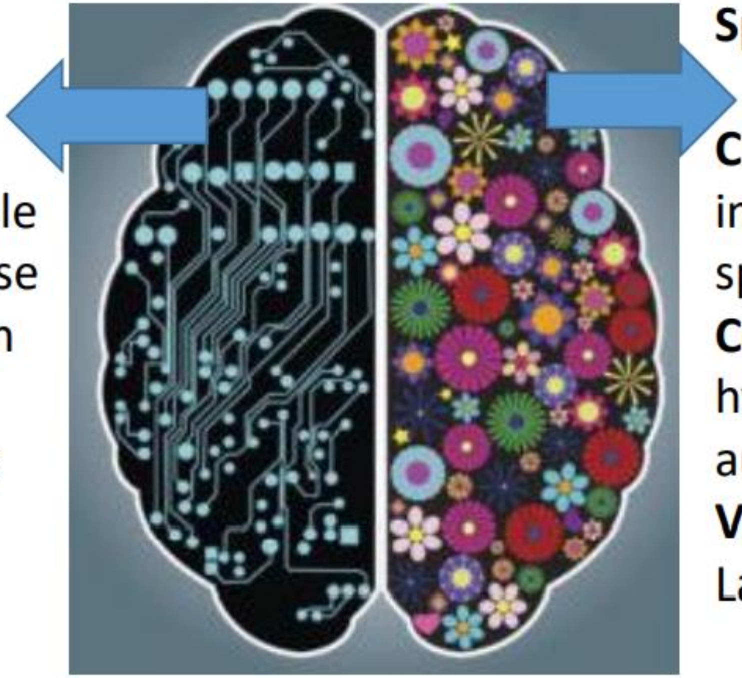


Image Sources: businessjournalism.org
health.harvard.edu

Spatial commonsense (images)

CSK-SNIFFER: Generating adversarial images for object detection using spatial commonsense [AAAI 2020]
Conceptual Captions: Cleaned, hypernymed image alt-text data for automatic image captions [ACL 2018]
VISIR: Visual and Semantic Image Label Refinement [WSDM 2018]

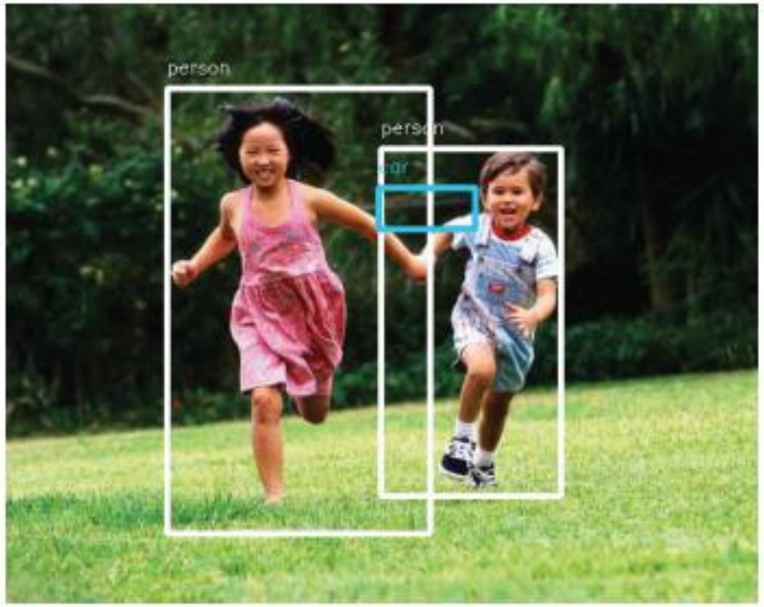
Extrinsic Evaluation

	Adversarial testing data	Test background knowledge
CSK-SNIFFER	Y	
Conceptual Captions		Y
VISIR		Y
WinoGrande	Y	
DoQA		Y
Arc	Y	

Extrinsic Evaluation

CSK-SNIFFER

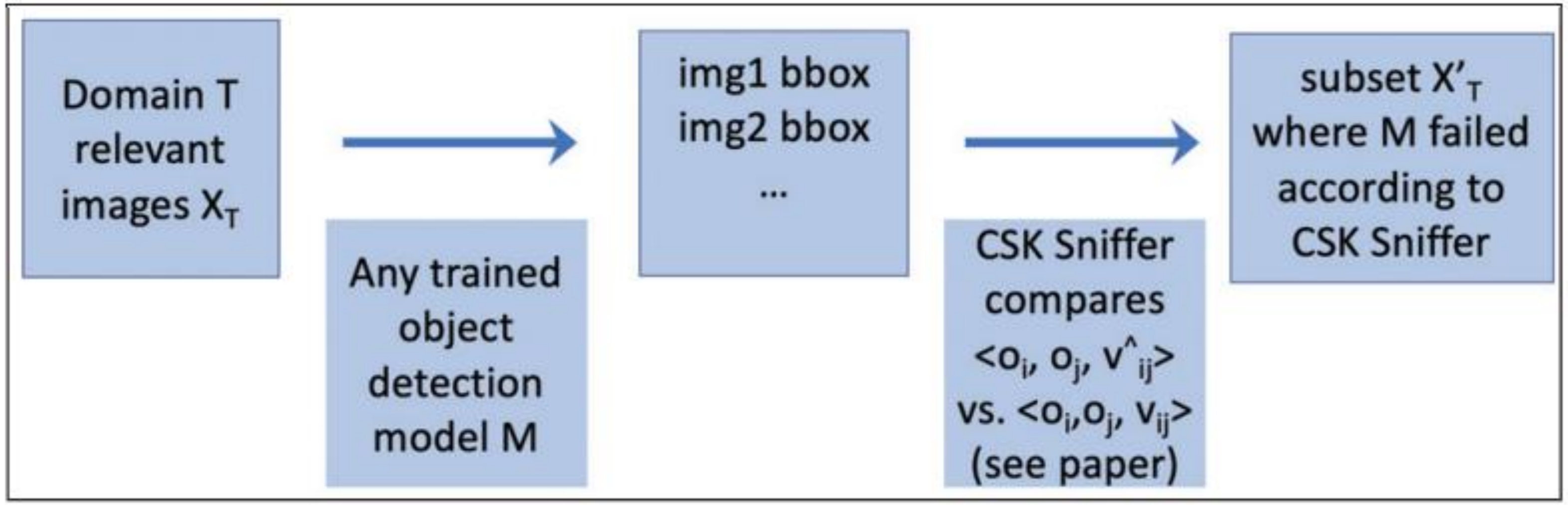
Spatial CommonSense



I Am Guessing You Can' t Recognize This: Generating Adversarial Images for Object Detection Using Spatial Commonsense (Student Abstract). AAAI. 2020

Extrinsic Evaluation

CSK-SNIFFER



Compare:
$< o_i, o_j, v_{ij} >$Two objects: o_i, o_j, relation: $v_{ij} \in rel(KB)$ [isAbove, isBelow, isInside, isNear, overlapsWith]

Thanks

20210526

Outline

- What is CommonSense Knowledge (CSK)
- Design Approach
- Extraction
- Consolidation
- Evaluation
- **Forward**

CSK Generation

Knowledge in Language Models

CSK Generation

Do language models have commonsense?

Prompts	
manual	DirectX *is developed by* y_{man}
mined	y_{mine} *released the* DirectX
paraphrased	DirectX *is created by* y_{para}

Top 5 predictions and log probabilities

	y_{man}		y_{mine}		y_{para}	
1	Intel	-1.06	Microsoft	-1.77	Microsoft	-2.23
2	Microsoft	-2.21	They	-2.43	Intel	-2.30
3	IBM	-2.76	It	-2.80	default	-2.96
4	Google	-3.40	Sega	-3.01	Apple	-3.44
5	Nokia	-3.58	Sony	-3.19	Google	-3.45

Jiang et al., TACL 2020

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

Feldman et al., EMNLP 2019

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

Weir et al., CogSci 2020

CSK Generation

Do language models have commonsense

- Distinction between **encoding** commonsense knowledge and **expressing** commonsense knowledge
- Probing with prompts measures whether LMs can **express** commonsense knowledge and the results are **mixed**

CSK Generation

Do Language Models know this

Sentence:

mango is a

Predictions:

2.1% great
1.9% very
1.2% new
1.0% good
1.0% small
← Undo

a mango is a

4.2% good
4.0% very
2.5% great
2.4% delicious
1.8% sweet
← Undo

Sentence:

A mango is a

Predictions:

4.2% fruit
3.5% very
2.5% sweet
2.2% good
1.5% delicious
← Undo

CSK Generation

Do Masked Language Models know this

Sentence:

mango is a [MASK]

Mask 1 Predictions:
69.7% .
9.3% ;
1.7% !
0.8% vegetable
0.7% ?

Sentence:

mango is a [MASK] .

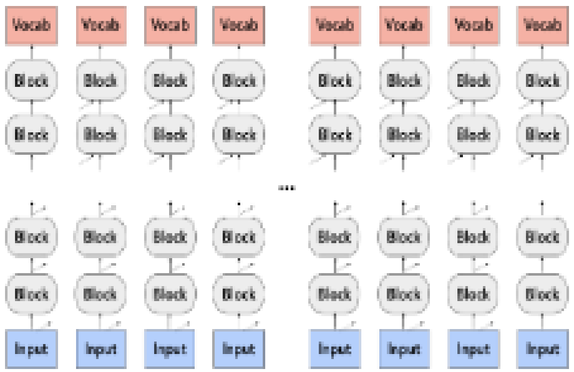
Mask 1 Predictions:
7.6% staple
7.6% vegetable
4.6% plant
3.5% tree
3.5% fruit

Sentence:

A mango is a [MASK] .

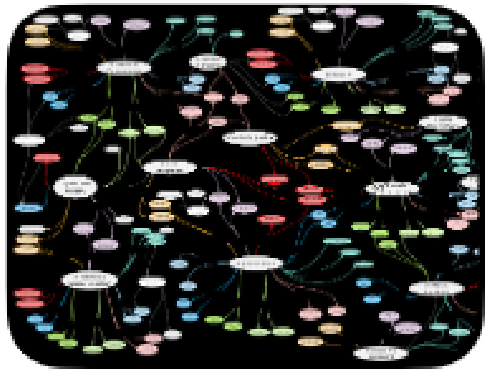
Mask 1 Predictions:
16.0% banana
12.1% fruit
5.9% plant
5.5% vegetable
2.5% candy

CSK Generation



Pre-trained
Language Model

+



Seed Knowledge
Graph Training

=

COMET

COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. ACL. 2019