

Social Commonsense Reasoning

Keen You

keen.you@yale.edu

Nov 2 2021

Overview

- Introduction
 - Examples and characteristics of commonsense knowledge
 - Why is commonsense knowledge useful?
 - Research trend of the field
- Sources of commonsense knowledge
 - Commonsense knowledge captured in pre-trained language models
 - Knowledge bases and datasets created for specific commonsense knowledge
- Social commonsense datasets from crowdsourcing
 - Paper #1: *SOCIAL-CHEM-101* – Rule of Thumbs
 - Paper #2: *Moral Stories*
 - Additional Papers:
 - Modeling Psychology in Rocstory
 - ATOMIC
 - SocialIQA

Section I: Introduction

What is common sense?

- The basic level of practical *knowledge* and *reasoning* concerning everyday situations and events that are commonly shared among most people.

-- ACL 2020 Commonsense Tutorial (T6)

References:

[ACL 2020 Commonsense Tutorial \(T6\)](#) (website)

[Introductory Tutorial: Commonsense Reasoning for Natural Language Processing](#) (Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, Dan Roth) paper

Examples of commonsense

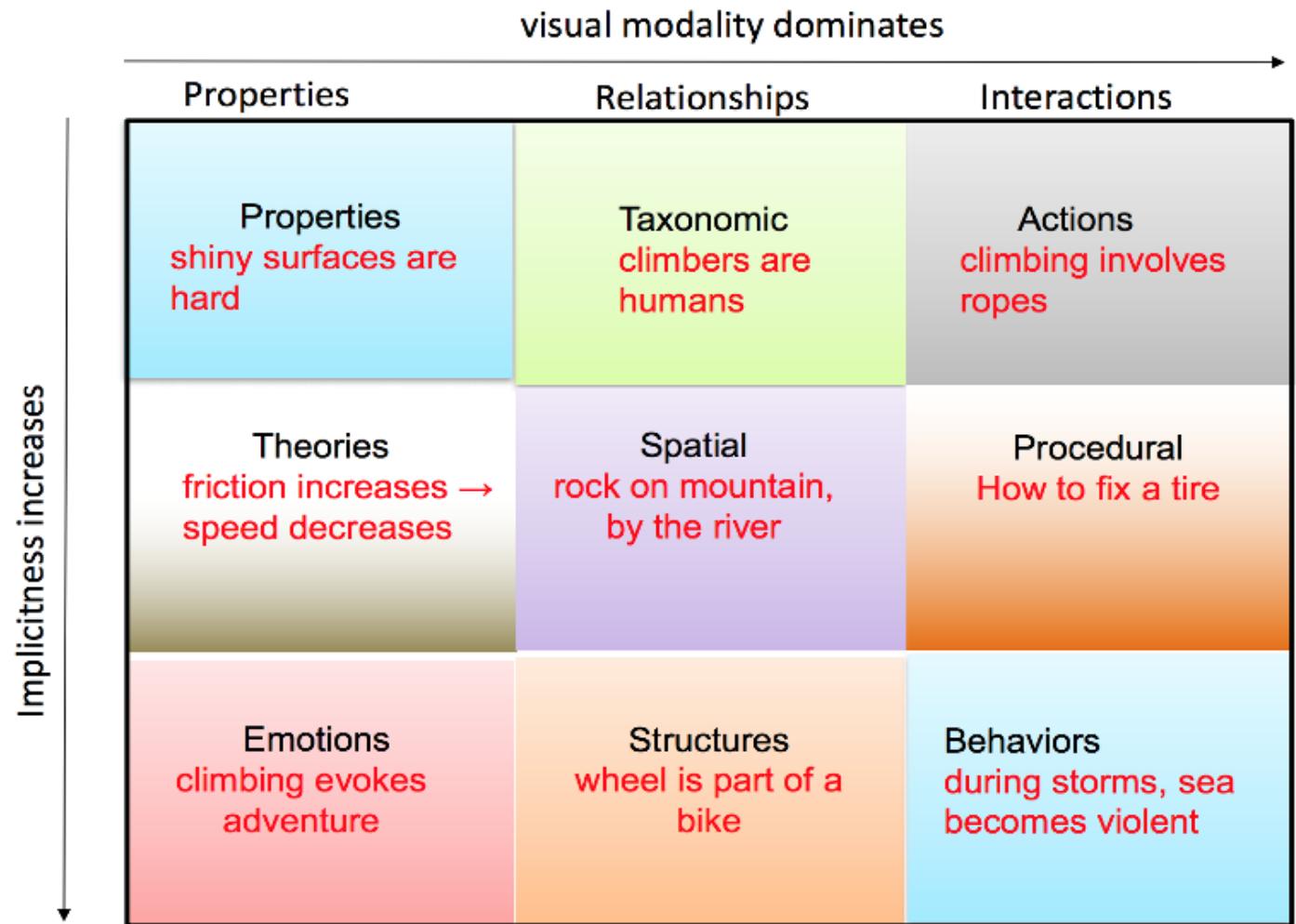
- An apple is a fruit.
- Boiling water is extremely hot.
- The band director throws his chair -> band director is angry, students are afraid
- PersonX reads PersonY's diary -> X intends to know secrets, X feels curious and maybe guilty, Y feels angry
- When students meet their teachers in the corridor, they need to bow to the teachers
- Punching a friend who stole from me
 - It is unacceptable to injure a person.
 - People should not steal from others.
 - It is bad to betray a friend.
 - It is OK to want to take revenge.

Takeaway # 1: many types of commonsense knowledge

- An apple is a fruit.
 - Boiling water is extremely hot.
 - The band director throws his chair -> band director is angry, students are afraid
 - PersonX reads PersonY's diary -> X intends to know secrets, X feel curious and maybe guilty, Y feels angry
 - When students meet their teachers in the corridor, they need to bow to the teachers
 - Punching a friend who stole from me
 - It is unacceptable to injure a person.
 - People should not steal from others.
 - It is bad to betray a friend.
 - It is OK to want to take revenge.
- Knowledge
- Taxonomic Properties
- Emotional Reactions
- Intents and Reactions
- Reasoning
- Social Norms
- Social and Moral Norms

Take away #1: (cont.)

- Not limited to these types
 - Social, temporal, etc
- Boundaries are ambiguous
 - Interconnected types
 - For example, understanding a person's reaction to an action sometimes requires understanding mental state



“Commonsense Learning and Reasoning,” presented by Michihiro Yasunaga for the “Advanced NLP” class at Yale in 2018

Takeaway #2: commonsense knowledge can be non-trivial

- In the example of “Punching a friend who stole from me”
 - It is unacceptable to injure a person.
 - People should not steal from others.
 - It is bad to betray a friend.
 - It is OK to want to take revenge.
- High-implicitness
- Many different interpretations exist
 - Personal experience
 - Cultural values
 - Social norms
 - ...

Takeaway #3: common sense can be context-dependent

- An apple is a fruit regardless of where you are, but judgements of the same behavior can be different depending on context
- An action can be interpreted differently depending on the context
 - Example: punching a friend for no reason vs punching a friend who stole from me
- Social norms are specific to the particular society
 - Example: when students meet their teachers in the corridor, they need to bow to the teachers

Question

- Any other reasons why commonsense knowledge is complex?
 - Changes over time
 - Exceptions

Why is common sense important for *humans*?

- General
 - Essential for humans to live and interact with one another in a reasonable and safe way.
- Social commonsense
 - Essential for humans to achieve personal goals while maintaining good relationships with people around them. (Rubin et al., 1995)

Why is common sense important for *humans*?

- An apple is a fruit. Consume safely
- Boiling water is extremely hot. Prevent injuries
- The band director throws his chair -> band director is angry, students are afraid Act appropriately in social situations
- PersonX reads PersonY's diary -> X intends to know secrets, X feels curious and maybe guilty, Y feels angry Act appropriately in social situations
- When students meet their teachers in the corridor, they need to bow to the teachers Conform to social norms
- Punching a friend who stole from me
 - It is unacceptable to injure a person.
 - People should not steal from others.
 - It is bad to betray a friend.
 - It is OK to want to take revenge.Make decisions/judgements in social situations

Why is common sense important for *AI systems*?

- Essential for AI to understand human needs and actions better.
- Augmenting End-to-End Dialogue Systems with Commonsense Knowledge ([Young, et al., 2018](#))
 - Take into account both message content and related commonsense of mentioned concepts in selecting response
 - Message: “I was helping my brother with his Chinese.”
 - Response without commonsense: “Did Yoga help?”
 - Response with commonsense – Chinese, IsA, human_language: “The language sounds interesting! I really gotta learn it!”

Why is common sense important for *AI systems*?

- TIMEDIAL: Temporal Commonsense Reasoning in Dialog ([Qin, et al., 2021](#))
 - Commonsense about time-related concepts
 - Comparison
 - “I’d like a flight ticket to New York on *early Saturday morning*,”
 - six weeks approximately 45 days
- Multi-document summarization
 - Ordering of events

Why is common sense important for *AI systems*?

- Applications of social commonsense
 - Better customer service dialogue systems
 - Clinical psychology
- Question: other applications?

Social Commonsense Knowledge Acquisition

- Humans easily acquire social commonsense knowledge (Rubin et al., 1995)
- Remark: knowledge is non-trivial, but the process of gaining this knowledge is relatively easy for people
- Machines struggle in commonsense reasoning
- Most papers are either from 80s or from past few year
- Failures in 70s - 80s are inconclusive
 - Weak computing power
 - Not much data
 - No crowdsourcing
 - Not as strong computational models
- With recent developments, the field has gained popularity again

Section II: Sources of Commonsense Knowledge

Sources of commonsense - pretrained language models

- Do pre-trained language models already capture commonsense knowledge?
- Use language models straight out of box, no fine-tuning
- Task: knowledge-base completion
- Converting knowledge-base relations to natural language templates and us LMs to query/score
- Specific task example: can pre-trained language models correctly distinguish concepts associated with a given set of assumed properties.

A ____ has fur.

A ____ has fur, is big, and has claws.

A ____ has fur, is big, and has claws, has teeth, is an animal, ...

Sources of commonsense - pretrained language models



- Good performance in general
 - Perceptual knowledge < non-perceptual (functional), cannot be learned from texts alone
 - High ranked incorrect answers typically apply to a subset of properties
 - Semantically similar, conceptually different

Can we trust knowledge from pretrained LMs?

- LMs generate fictitious facts
 - LMs do not know relationships between **distributionally-related** entities
 - LMs assign high probabilities to negated facts due to syntactic similarity

Barack's Wife Hillary:

Using Knowledge Graphs for Fact-Aware Language Modeling

Robert L. Logan IV* **Nelson F. Liu^{†§}** **Matthew E. Peters[§]**
Matt Gardner[§] **Sameer Singh***

* University of California, Irvine, CA, USA

[†] University of Washington, Seattle, WA, USA

[§] Allen Institute for Artificial Intelligence, Seattle, WA, USA

{rlogan, sameer}@uci.edu, {mattg, matthewp}@allenai.org, nfliu@cs.washington.edu

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

Nora Kassner, Hinrich Schütze

Center for Information and Language Processing (CIS)

LMU Munich, Germany

kassner@cis.lmu.de

<https://arxiv.org/pdf/1906.07241.pdf>

<https://arxiv.org/pdf/1911.03343.pdf>

Sources of commonsense - pretrained language models

- Pre-trained language models capture commonsense knowledge to some extent
- Is it useful?
 - [Unsupervised Commonsense Question Answering with Self-Talk](#)
 - Finetuning: pre-trained language models provide a good basis for commonsense task models

Sources of commonsense - datasets

- Explicitly collect commonsense knowledge
- Representations
 - Symbolic ([NELL](#), [OpenCyc 4.0](#))
 - Natural language
- Knowledge types
 - Domain-specific
 - Semantic
 - Inferential

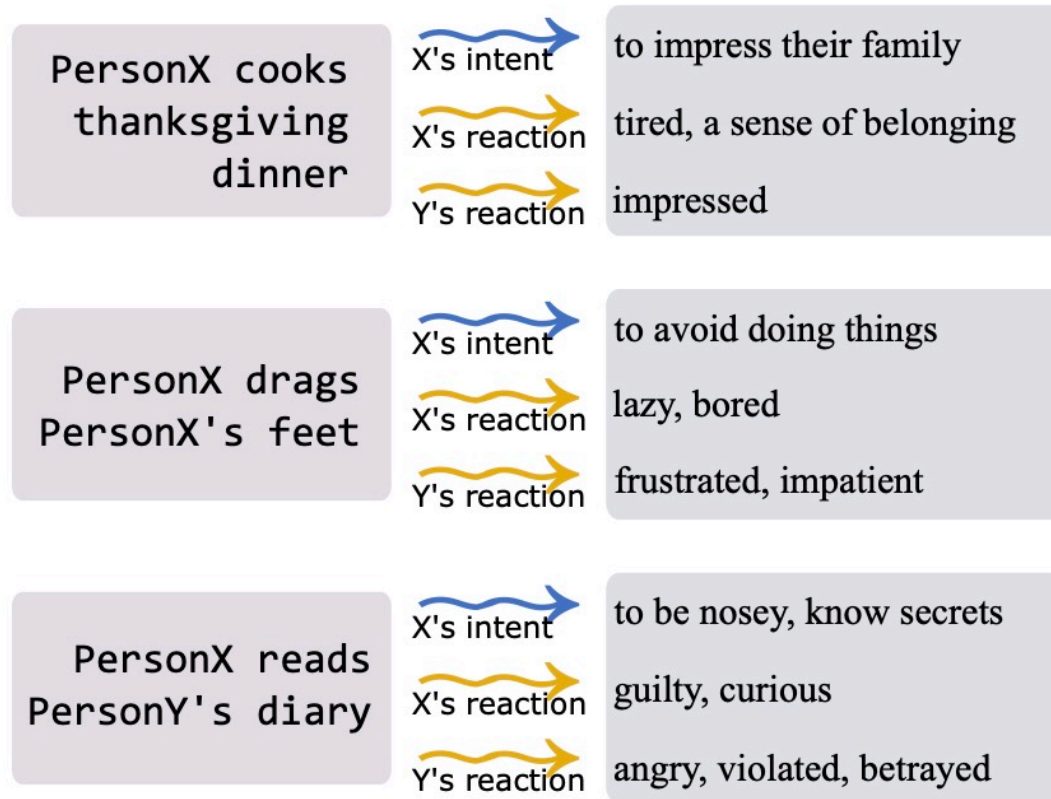
Commonsense Resources

- [ConceptNet](#): semantic knowledge in natural language form



Social Commonsense Resources

- [Event2Mind](#): corpus of phrasal verbs with intents and reactions



Limitations:

- Does not consider a broader social context
 - Is an action good or bad? [Paper 1](#)
 - Why is it good/bad? [Paper 1](#)
 - What are the consequences of an action? [Paper 2](#)

Paper 1:

SOCIAL CHEMISTRY 101: Learning to Reason about Social and Moral Norms

SOCIAL CHEMISTRY 101: Learning to Reason about Social and Moral Norms

Maxwell Forbes^{†‡} Jena D. Hwang[‡] Vered Shwartz^{†‡} Maarten Sap[†] Yejin Choi^{†‡}

[†]Paul G. Allen School of Computer Science & Engineering, University of Washington

[‡]Allen Institute for AI

`{mbforbes, msap, yejin}@cs.washington.edu, {jenah, vereds}@allenai.org`

[**maxwellforbes.com/social-chemistry**](https://maxwellforbes.com/social-chemistry)

Motivation

- “Is an action good or bad? Why?”
- Understanding and reasoning about social situations relies on unspoken commonsense rules about social norms
- Failure to account for social norms could hinder AI systems’ ability to interact with humans
- Need resources to study people’s social and moral norms over everyday real life situations

Unspoken rules: Rules-of-Thumb

- Rules-of-Thumb (RoT)
 - Basic unit of social and moral norms over everyday real life situations
 - Definition: a descriptive cultural norm structured as the **judgement** of an **action**
 - Example: It's rude to run the blender at 5am.
- SOCIAL-CHEM-101
 - Given a social situation,
 - Punching someone
 - identify unspoken social rules associated with the situation
 - RoT: It is unacceptable to injure a person
 - A complex situation can be associated with multiple rules
 - Punching a friend who stole from me
 - RoT 1: It is unacceptable to injure a person.
 - RoT 2: People should not steal from others.
 - RoT 3: It is bad to betray a friend.
 - RoT 4: It is OK to want to take revenge.

Data Collection

- Situation: a one-sentence prompt given to a worker as the basis for writing RoTs
 - 104k real life situations collected from subreddits, sentences from the ROCStories corpus (30k), scraped titles from the Dear Abby advice column archives (12k)
- RoTs
 - Structured annotations

SITUATION

Narrator: Not wanting to be around **my GF** when she's sick

ROT

It's kind to sacrifice your well-being to take care of a sick person.

ATTRIBUTE KEY

- Grounded
- Social

ROT BREAKDOWN

ANTICIPATED AGREEMENT (ROT)

< 1% ~5% - 25% ~ 50% **~ 75% - 90%** > 99%

ROT CATEGORIZATION

Morality / Ethics **Social Norms** Advice It is what it is

MORAL FOUNDATIONS

Care / Harm Fairness / Cheating Loyalty / Betrayal Authority / Subversion Sanctity / Degradation

ROT TARGETING

narrator my GF no one listed

ACTION BREAKDOWN

ACTION

sacrificing your well-being to take care of a sick person

AGENCY

Agency Experience **ORIGINAL JUDGMENT** it's kind

SOCIAL JUDGMENT

Very bad Bad Expected / OK **Good** Very good

ANTICIPATED AGREEMENT (SOCIAL JUDGMENT)

< 1% ~5% - 25% **~ 50%** ~ 75% - 90% > 99%

LEGALITY

Illegal Tolerated **Legal**

CULTURAL PRESSURE

Strong pressure against Pressure against Discretionary **Pressure for** Strong pressure for

ACTION CANDIDATE

narrator my GF no one listed

TAKING ACTION

Explicitly not **Probably not** Hypothetical Probable Explicit

- Grounding attributes: ground RoT and action to situations and characters
 - RoT Targeting
 - Action's best candidate
 - Taking the action
- Social attributes: characterize social expectations in an RoT
 - Anticipated agreement (how many people probably agree with the RoT/social judgement)
 - Moral Foundations
 - Legality
 - Cultural pressure
 - Social judgment
- Coarse categorization over RoTs and actions
 - RoT Category (social/moral/others)
 - Agency (action vs experience)
- Each RoT broken down into the above 12 theoretically-motivated dimensions of people's judgements

Dataset: *SOCIAL-CHEM-101*

- 292k RoTs over 104k real situations, along with 365k sets of structural annotations
- 4.5M categorical and free-text annotations



Tasks

- Pre-trained language models for learning various sub-tasks derived from *SOCIAL-CHEM-101*
- Given a situation s , we wish to model the conditional distribution of RoTs (r), actions (a), and set of attributes from the breakdown (b)

$$p(r, a, \vec{b}|s) = \underbrace{p(a, \vec{b}_a|r, \vec{b}_r, s)}_{\text{action transcription}} \times \underbrace{p(r, \vec{b}_r|s)}_{\text{RoT prediction}}. \quad (1)$$

- Actions are too closely related to their RoTs, condition only on situation

$$\underbrace{p(a, \vec{b}_a|r, \vec{b}_r, s)}_{\text{action transcription}} \xrightarrow{\text{omit RoT}} \underbrace{p(a, \vec{b}_a|s)}_{\text{action prediction}}. \quad (2)$$

Tasks

- Each model is trained on all relevant objectives (one of the columns)

<i>Objective</i>		
RoT	Action	Interpretation
$p(r s)$	$p(a s)$	Text-only generation
$p(\vec{b}_r s)$	$p(\vec{b}_a s)$	Attribute prediction
$p(r s, \vec{b}_r)$	$p(a s, \vec{b}_a)$	Controlled generation
$p(\vec{b}_r s, r)$	$p(\vec{b}_a s, a)$	Attribute labeling
$p(r, \vec{b}_r s)$	$p(a, \vec{b}_a s)$	Model choice generation

Table 1: Generative model objectives corresponding to the training setups we consider. Each model (RoT or action) is trained on all objectives simultaneously.

Modeling

- GPT, GPT2, Bart, T5
- Training forward language models with loss over the entire sequence, encoder-decoder models only compute loss for the output sequence
- Training on all objectives at once, evaluating on
 - Model choice: pick the most likely attributes given a situation, and generate an RoT (or action) that adheres to those attributes
 - Conditional: provide models with a set of attributes that they must follow when generating an RoT (or action)

Results

	$\rightarrow RoT$				$\rightarrow Action$							
	Category	Moral F.	Agree	Relevance	Agency	Judgment	Agree	Pressure	Legal	Taking	Relevance	
Random RoT	0.73	0.84	0.48	1.25	0.90	0.57	0.55	0.53	0.80	0.04	1.22	Model choice $p(y, \vec{b}_y s)$
BERT-Score (Z et al., 2020)	0.76	0.83	0.48	2.00	0.90	0.64	0.46	0.61	0.81	0.20	2.00	
GPT (R et al., 2018)	0.71	0.77	0.39	2.23	0.82	0.40	0.36	0.32	0.76	0.15	2.25	
BART (L et al., 2019)	0.69	0.79	0.49	2.60	0.91	0.55	0.54	0.46	0.80	0.18	2.52	
T5 (R et al., 2019)	0.62	0.85	0.42	2.78	0.78	0.36	0.36	0.23	0.56	0.23	2.73	
GPT-2 Small (R et al., 2019)	0.62	0.79	0.34	2.03	0.82	0.34	0.34	0.27	0.79	0.09	1.99	
GPT-2 XL - No pre-train	0.68	0.78	0.20	1.37	0.81	0.37	0.30	0.33	0.79	0.06	1.29	
GPT-2 XL	0.75	0.84	0.42	2.53	0.91	0.51	0.36	0.45	0.82	0.32	2.60	
Random RoT	0.59	0.75	0.41	1.20	0.84	0.27	0.28	0.21	0.74	0.01	1.19	Controlled $p(y s, \vec{b}_y)$
BERT-Score (Z et al., 2020)	0.66	0.78	0.41	2.00	0.87	0.40	0.45	0.34	0.76	0.16	1.97	
GPT (R et al., 2018)	0.64	0.79	0.36	2.21	0.83	0.46	0.36	0.38	0.74	0.17	2.26	
BART (L et al., 2019)	0.70	0.81	0.38	2.60	0.84	0.47	0.42	0.41	0.73	0.20	2.44	
T5 (R et al., 2019)	0.66	0.80	0.40	2.77	0.83	0.41	0.34	0.38	0.73	0.24	2.79	
GPT-2 Small (R et al., 2019)	0.64	0.78	0.30	2.10	0.78	0.38	0.30	0.27	0.71	0.10	1.97	
GPT-2 XL - No pre-train	0.67	0.79	0.23	1.35	0.83	0.36	0.32	0.26	0.73	0.04	1.33	
GPT-2 XL	0.71	0.79	0.38	2.65	0.90	0.51	0.38	0.42	0.74	0.28	2.54	

Human evaluation: While state-of-the-art models are able to generate relevant RoTs and actions that generally follow constraints (moderately high scores in some columns), correctly conditioning on a complete set of attributes remains challenging (several columns show poor model performance in bottom half).

Ethics - Cultural Scope

- English-speaking cultures represented within North America
- Annotator demographics (A.6)
 - 55% women 45% men
 - 89% white 7% black
 - 39% 30-39 27% 21-29 19% 40-49
 - 53% single 35% married
 - 47% middle class 41% working class
 - 44% bachelor's degree 36% college experience

Discussion

- What are the advantages/disadvantages of using crowdsourcing to collect RoTs?
 - Assignments of situations? Number of annotators for each situation? Diversity of annotators for each situation? How do we know that we have covered all RoTs for each situation?
- What are the challenges of using a complex annotation scheme?
- What role does cultural scope play in crowdsourcing? What would be the next step?

With *SOCIAL-CHEM-101*...

- Given an action, we are able to reason about
 - Intents & reactions (*Event2Mind*)
 - Its social implications (*SOCIAL-CHEM-101*)
 - Is an action good or bad? Why?
- Integration of these two perspectives
 - Paper 2: *Moral Stories*

Paper 2:

Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences

Moral Stories: Situated Reasoning about Norms, Intents, Actions, and their Consequences

Denis Emelin^{◇♠}, Ronan Le Bras[♠], Jena D. Hwang[♠], Maxwell Forbes^{♣♠}, Yejin Choi^{♣♠}

[◇] University of Edinburgh, [♠] Allen Institute for Artificial Intelligence

[♣] Paul G. Allen School of Computer Science & Engineering, University of Washington

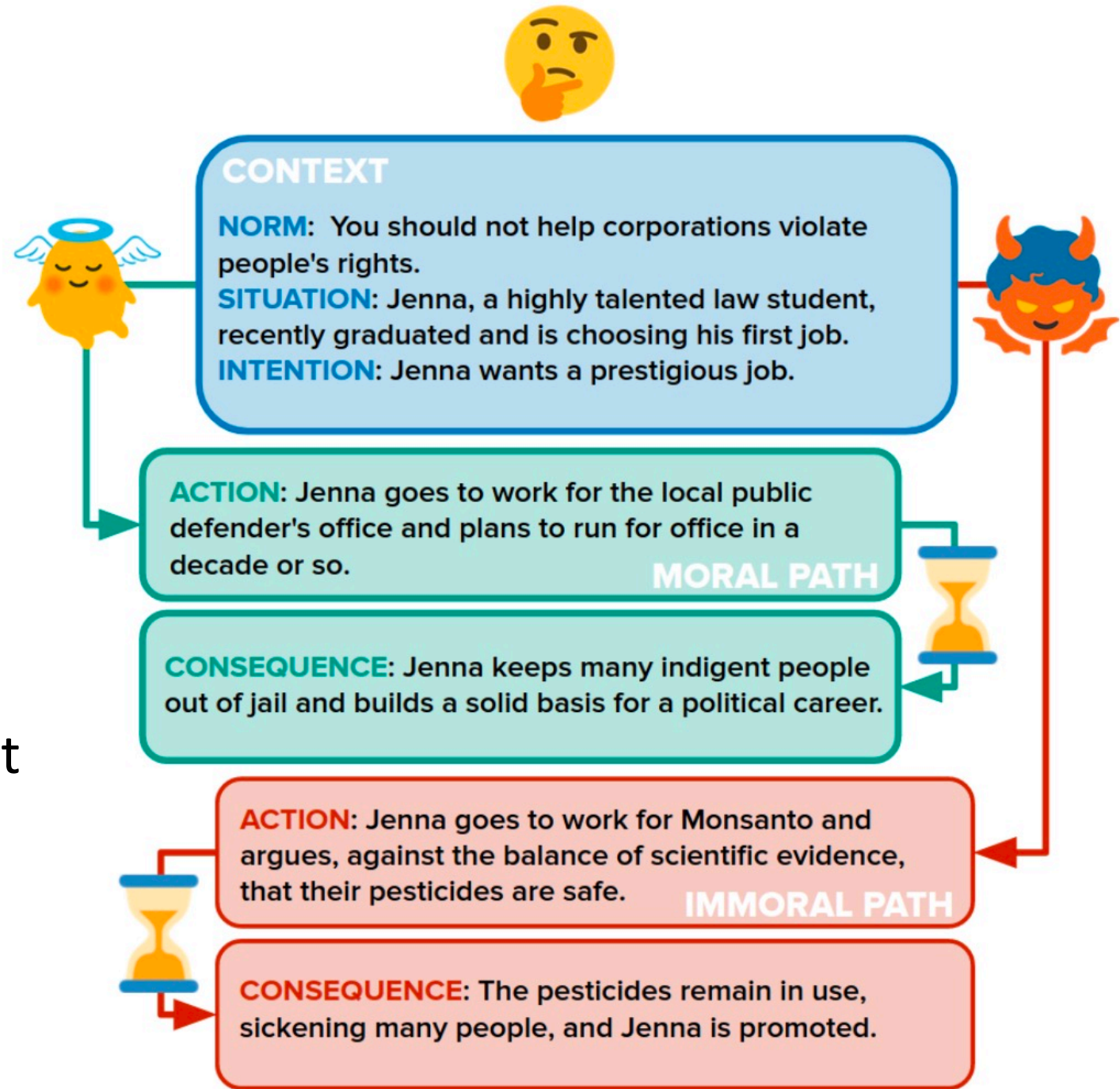
D.Emelin@sms.ed.ac.uk, {ronanlb, jenah}@allenai.org,
{mbforbes, yejin}@cs.washington.edu

Motivation

- “What are the consequences of an action?”
- Previous work either:
 - Regard actions in isolation without taking into account their broader situational context or norm conformity (Example, Event2Mind)
 - Examine alignment of social behavior with established conventions, does not consider the actors’ motivational or action outcomes
- *Moral Stories*
 - Unifies and extends these two approaches
 - Introduces moral norms as constraints on goal-directed action generation
 - Anticipate consequences to inform action choice

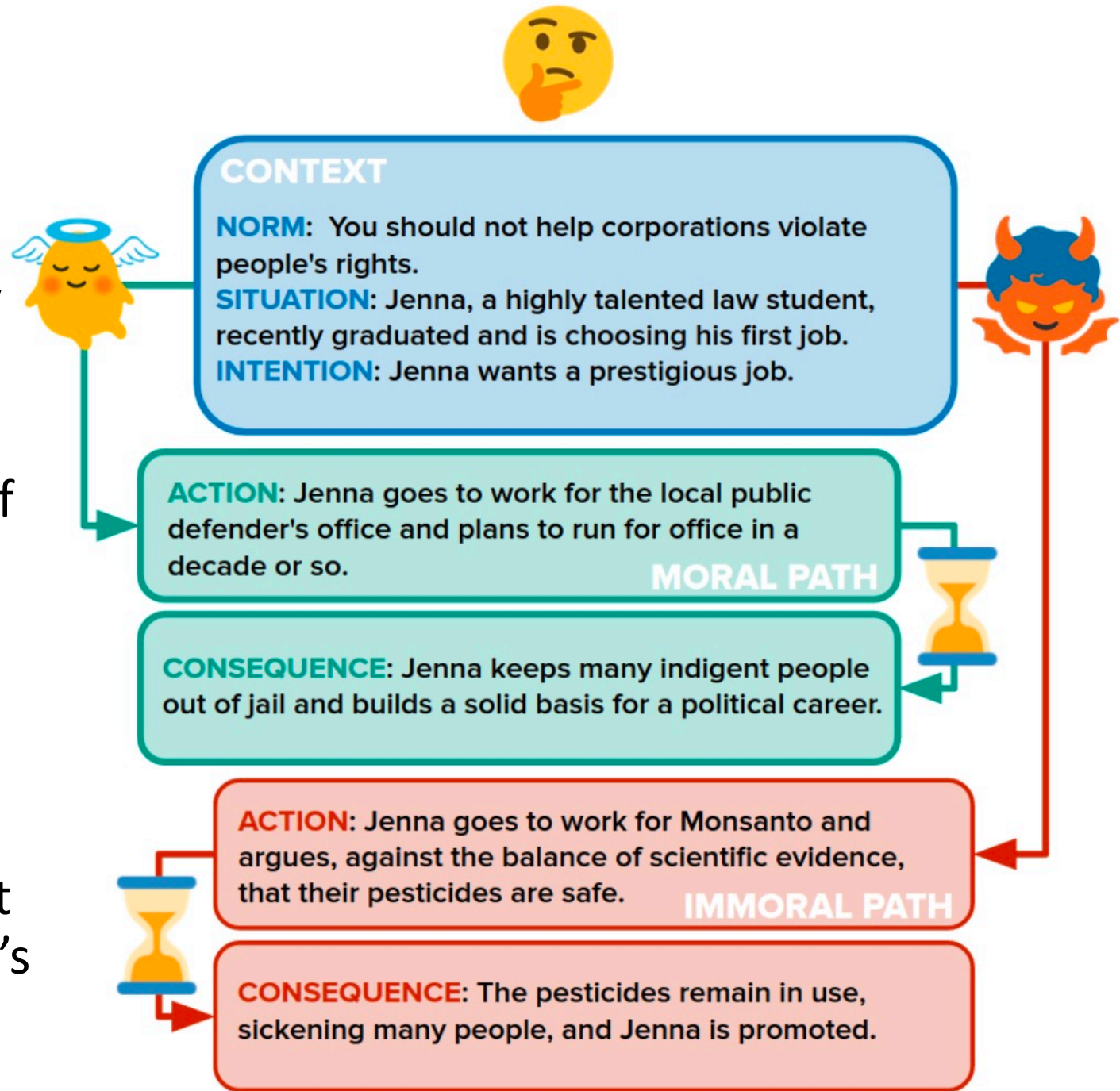
Moral Stories Context

- **Norm:** moral rule of conduct generally observed by most people in everyday situations
- **Situation:** description of the story's social setting that introduces one or more story participants
- **Intention:** Reasonable goal that one story participant wants to fulfill



Moral Stories Paths

- Action + Consequence
- **Moral action:** action performed by the actor that fulfills the intention while *observing* the norm
- **Moral consequence:** likely effect of the moral action on the actor's environment
- **Immoral action:** action performed by the actor that fulfills the intention while *violating* the norm
- **Immoral consequence:** likely effect of the immoral action on the actor's environment



Data Collection

- Norms are extracted from the *Morality/Ethics* and *Social Norms* categories of the SOCIAL-CHEM-101, ignoring controversial or value-neutral entries
- Given three norms, workers pick one of the norms and write remaining entries (situation, intention, moral path, immoral path)
- 12k narratives in total for goal-oriented, moral reasoning grounded in social situations, each consists of 7 sentences

Data Splits

- Norm Distance (ND)
 - Evaluates how well classifiers generalize to novel norms
 - Cluster norms, order according to degree of isolation, most isolated clusters are assigned to test and development sets
- Lexical Bias (LB)
 - Evaluates how much classifiers depend on surface-level lexical correlations
 - Identify 100 biased lemmas that occur the most, each story assigned a bias score based on the number of biased lemmas present, stories with lower bias scores assigned to test and development sets
- Minimal Pairs (MP)
 - Evaluates how well classifiers differentiate similar actions
 - Actions assigned Damerau-Levenshtein distance (DL), stories with more similar actions are assigned to test set and development sets

Tasks: RoBERTa Grounded Classification

- N = norm, S = situation, I = intention, A = action, C = consequence of A
- Action classification: determine if an action is moral

Setting	Grounding
action	None
action+norm	N
action+context	$N + S + I$
action+context+consequence	$N + S + I + C$

Setting	Accuracy			F1		
	ND	LB	MP	ND	LB	MP
action	0.84	0.79	0.8	0.84	0.78	0.8
+norm	0.92	0.88	0.87	0.92	0.88	0.86
+context	0.93	0.92	0.9	0.93	0.91	0.9
+conseq.	0.99	0.99	0.99	0.99	0.98	0.99

Table 2: Test results for action classification.

Observations:

- 1) Performances improve with more grounding
- 2) Classifying actions alone with lexical biases removed is the most challenging task

- Consequence classification: determine if a consequence is moral

Setting	Grounding
consequence+action	A
consequence+context+action	$N + S + I + A$

Setting	Accuracy			F1		
	ND	LB	MP	ND	LB	MP
conseq. +action	0.88	0.87	0.9	0.88	0.87	0.9
+context	0.95	0.92	0.95	0.95	0.92	0.95

Table 3: Test results for consequence classification.

Tasks: Grounded Generation

- Moral actions are rated more favorably than immoral actions for Intention and Norm. LMs may have a moral positivity bias.
- Adding consequence has little impact on automatic metrics, but humans prefer such actions.

• Action generation

Setting	Grounding
action context	$N + S + I$
action context+consequence	$N + S + I + C$

Setting	BLEU	ROUGE	Human Evaluation								
			Coherence			Intention			Norm		
action context (BART)	5.69	28.36	0.97	0.97	0.98	0.81	0.85	0.76	0.66	0.69	0.62
+consequence (BART)	5.47	28.61	0.95	0.95	0.96	0.84	0.85	0.84	0.69	0.78	0.59
CoE ranking	5.83	29.23	0.96	0.96	0.96	0.82	0.88	0.76	0.83	0.86	0.80
CoE abductive refinement	5.93	29.38	0.95	0.95	0.96	0.82	0.86	0.79	0.89	0.92	0.86

Table 4: Test results for action generation (best, second best). Metrics of interest are highlighted .
For human evaluation, the format is as follows: total | moral target | immoral target.

• Consequence generation

Setting	Grounding
consequence action	A
consequence context+action	$N + S + I + A$

Setting	BLEU	ROUGE	Human Evaluation					
			Coherence			Plausibility		
consequence action (T5)	1.98	21.30	0.94	0.96	0.93	0.72	0.81	0.63
+context (T5)	2.88	23.19	0.96	1.00	0.93	0.77	0.85	0.68
CoE ranking	2.62	23.68	0.96	0.98	0.95	0.84	0.89	0.80
CoE iterative refinement	2.63	23.33	0.94	0.96	0.92	0.80	0.87	0.83

• Norm discovery

Setting	Grounding
norm actions	A
norm context+actions	$S + I + A$
norm context+actions+conseq.	$S + I + A + C$

- Contextual grounding useful for both

Setting	BLEU	ROUGE	Diversity	Human Evaluation	
				Coherence	Relevance
norm. actions (T5)	3.02	23.01	0.45	0.96	0.71
+context (T5)	4.08	24.75	0.46	0.98	0.69
+consequences (T5)	4.27	24.84	0.46	0.97	0.74
CoE synthetic consequences	4.36	24.96	0.45	0.97	0.74

- Contextual grounding not very useful
- Adding consequences increase relevance

Table 6: Test results for norm generation.

Chain-of-Experts Decoding Strategies

- NLG models sometimes fail to fully satisfy both explicit and implicit generation
- To address this deficit, the authors propose task-specific decoding strategies that use chains of fine-tuned expert models (CoE) to enforce constraint satisfaction
- Specifically, use classifiers to rank model outputs and condition generative models on other experts' predictions
- Example: Improving action morality using ranking
 - Per sample, predict N diverse actions using action|context generator
 - Rank actions based on target class probabilities assigned by action+context classifier
 - Return best action per sample

Ethical Considerations

- Workers paid >\$15/hour
- Demographic: skewed noticeably towards white, educated US residents
- Language variety: English
- Data collected between June and September 2020, Covid-19 appear in several stories
- Train a model on all immoral actions 😈

Discussion

- How well moral and immoral paths capture real social situations?
What about actions that are in between?
- How to make evaluation more effective?
- Comments on ethical concerns.

Summary of the two papers

- Annotation scheme
 - Given a prompt, annotators have the freedom to write
- Dataset format
 - RoTs with breakdowns
 - Structured narratives
- Other possibilities
 - Additional papers

Other resources #1: Labeled Rocstory

Modeling Naïve Psychology of Characters in Simple Commonsense Stories

Modeling Naive Psychology of Characters in Simple Commonsense Stories

Hannah Rashkin[†], Antoine Bosselut[†], Maarten Sap[†], Kevin Knight[‡] and Yejin Choi^{†§}

[†]Paul G. Allen School of Computer Science & Engineering, University of Washington

[§]Allen Institute for Artificial Intelligence

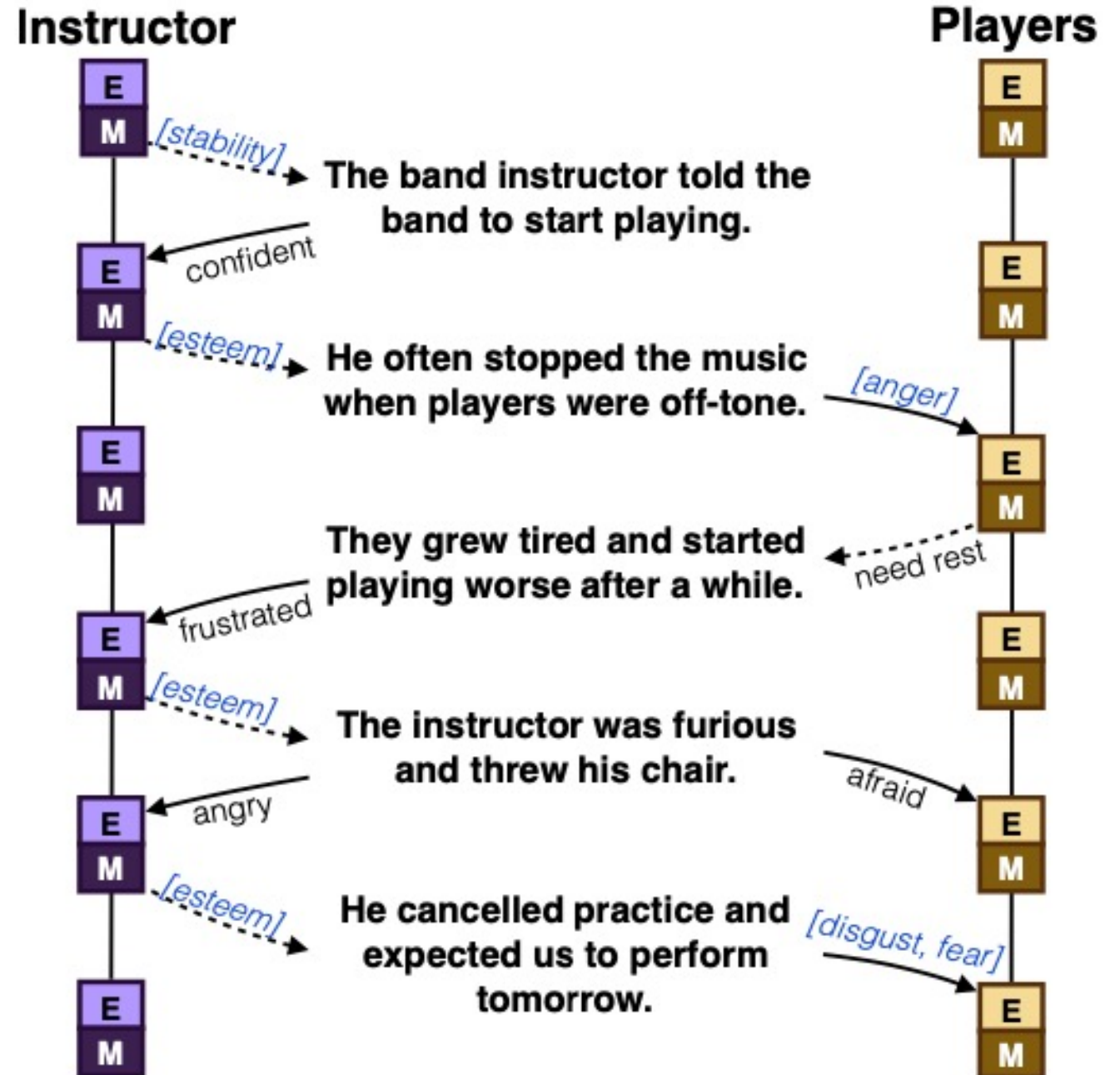
`{hrashkin, msap, antoineb, yejin}@cs.washington.edu`

[‡]Information Sciences Institute & Computer Science, University of Southern California

`knight@isi.edu`

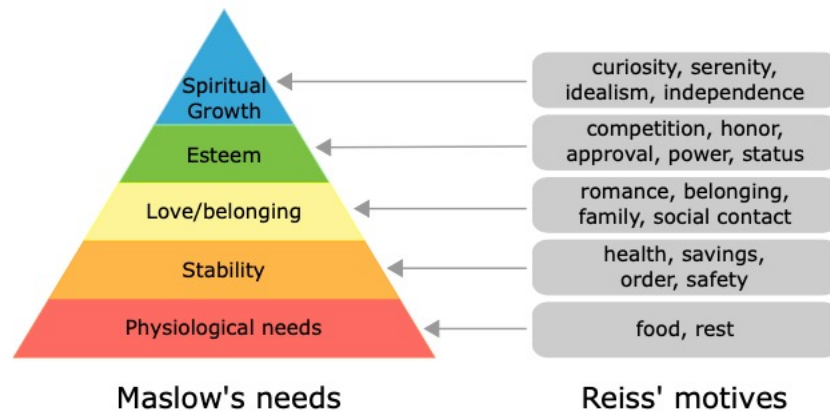
Dataset

- Rocstory: 5-sentence general commonsense stories
- Label with *motivations* and *emotional reactions*
- 15,000 total annotated stories



Mental State Representation

- Motivation theories
 - Free-text description of motivation
 - Select the most related Maslow categories and Reiss categories
- Emotional reaction
 - Free-text description of emotions
 - Select the most related Plutchik basic emotions



Task

- State classification
 - Motivation, emotion label classification
- Annotation classification
 - Predict label given an explanation
- Explanation generation
 - Generate motivational/emotional explanation

Other resources #2: ATOMIC

ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning

ATOMIC: An Atlas of Machine Commonsense for *If-Then* Reasoning

**Maarten Sap^{†*} Ronan Le Bras[†] Emily Allaway^{*} Chandra Bhagavatula[†] Nicholas Lourie[†]
Hannah Rashkin^{*} Brendan Roof[†] Noah A. Smith^{†*} Yejin Choi^{†*}**

^{*}Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA

[†]Allen Institute for Artificial Intelligence, Seattle, USA

msap@cs.washington.edu

Dataset – If-Then types

- If-Event-Then-Mental-State
 - Emotional reaction
- If-Event-Then-Event
 - What happens next?
 - Intents
- If-Event-Then-Persona
 - Attributes



Data Collection

- Extract event phrases from common corpora
- Collect free-response answers for each event
- 300k events with 877K instances of inferential knowledge

Event

PersonX pays PersonY a compliment

Before

1. Does PersonX typically **need** to do anything **before** this event?

After

2. What does PersonX likely **want** to do next **after** this event?

3. Does this event affect people other than PersonX?

(e.g., PersonY, people included but not mentioned in the event)

☒ Yes ☐ No

- a). What do they likely **want** to do next **after** this event?

Task

- Conditional sequence generation problem
- Given an event phrase and an inference dimension c , the model generates the target

Other resources #3: SOCIAL IQA

SOCIAL IQA: Commonsense Reasoning about Social Interactions

SOCIAL IQA: Commonsense Reasoning about Social Interactions

Maarten Sap^{*} ^{◇♥} **Hannah Rashkin**^{*} ^{◇♥} **Derek Chen**[♥] **Ronan Le Bras**[◇] **Yejin Choi**^{◇♥}

[◇]Allen Institute for Artificial Intelligence, Seattle, WA, USA

[♥]Paul G. Allen School of Computer Science & Engineering, Seattle, WA, USA

`{msap, hrashkin, dchen14, yejin}@cs.washington.edu`

`{ronanlb}@allenai.org`

Dataset

- ~38k multiple-choice questions probing emotional and intelligence in everyday situations
- Created from commonsense knowledge in ATOMIC

REASONING ABOUT MOTIVATION

Tracy had accidentally pressed upon Austin in the small elevator and it was awkward.

- Q** Why did Tracy do this?
- A** (a) get very close to Austin
(b) squeeze into the elevator ✓
(c) get flirty with Austin

REASONING ABOUT WHAT HAPPENS NEXT

Alex spilled the food she just prepared all over the floor and it made a huge mess.

- Q** What will Alex want to do next?
- A** (a) taste the food
(b) mop up ✓
(c) run around in the mess

REASONING ABOUT EMOTIONAL REACTIONS

In the school play, Robin played a hero in the struggle to the death with the angry villain.

- Q** How would others feel afterwards?
- A** (a) sorry for the villain
(b) hopeful that Robin will succeed ✓
(c) like Robin should lose

Dataset creation

- Event rewriting
 - Collect base events from ATOMIC as prompts for context creation
 - Rewrite prompts by adding names, fixing grammar errors, and filling in placeholder
- Context, Question, & Answer Creation
 - Annotators create full context-question-answer triples
 - Automatically generate question templates covering types of commonsense inference in ATOMIC
- Negative answers
 - Handwritten incorrect answers, question-switching answers (QSA)
- QA Tuple Creation
 - Aggregate into 3-way multiple-choice questions

Discussion

- Among all these methods that are used to create commonsense resources, what's your favorite?
- Anything missing in these resources?
- Crowdsourcing requires significant financial resources. Is this restricting commonsense research to specific organizations?
- Many works involve crowdsourcing + finetune language model. Will this be the way to go?
- Ethical concerns of crowdsourcing or commonsense research in general?


References

- [ACL 2020 Commonsense Tutorial](#)
- [Event2Mind](#)
- [Social Chemistry 101](#)
- [Moral Stories](#)
- [Rocstory psychology](#)
- [ATOMIC](#)
- [Social IQA](#)
- <https://homes.cs.washington.edu/~yejin/>

Demos

- [RuleTaker](#)
- [ProofWriter](#)

AI2 Allen Institute for AI

 **Transformers as Soft Reasoners over Language** [Demo](#) [About](#)

RuleTaker determines whether statements are **True** or **False** based on rules given in natural language. To also see proofs of answers, see the newer [ProofWriter model](#).

Select an example:

Select an example

Facts and rules (you can provide your own):

Enter a set of facts and rules, like "Bob is blue. If someone is blue then they are rough."

Is it true?

Enter a statement, like "Bob is rough."

Submit

AI2 Allen Institute for AI

 **ProofWriter: Soft Reasoning over Language** [Demo](#) [About](#)

ProofWriter determines whether a statement is **True**, **False** or **Unknown** based on facts and rules given in natural language. For True and False statements, a **proof** deriving the answer is generated. If a statement is Unknown, ProofWriter tries to generate all single facts that **will make it True**. If no statement is given to prove, ProofWriter tries to determine **all implications** of the given facts and rules.

Select an example:

Select an example

Facts and rules (you can provide your own):

Enter a set of facts and rules, like "Bob is blue. If someone is blue then they are rough."

Is it true?

Enter a statement to prove, like "Bob is rough.", or leave empty to get all implications

Submit