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

Maarten Sap[†⋆] **Ronan LeBras**[†] **Emily Allaway**[⋆] **Chandra Bhagavatula**[†] **Nicholas Lourie**[†]
Hannah Rashkin[⋆] **Brendan Roof**[†] **Noah Smith**[†⋆] **Yejin Choi**[†⋆]
[⋆]Paul G. Allen School for Computer Science and Engineering, University of Washington, Seattle
[†]Allen Institute for Artificial Intelligence, Seattle

Abstract

We present ATOMIC, an atlas of everyday commonsense reasoning, organized through 300k textual descriptions. Compared to existing resources that center around taxonomic knowledge, ATOMIC focuses on inferential knowledge organized as typed *if-then* relations with variables (e.g., "*if* X pays Y a compliment, *then* Y will likely return the compliment"). We propose nine *if-then* relation types to distinguish causes v.s. effects, agents v.s. themes, voluntary v.s. involuntary events, and actions v.s. mental states. By generatively training on the rich inferential knowledge described in ATOMIC, we show that neural models can acquire simple commonsense capabilities and reason about previously unseen events. Experimental results demonstrate that multitask models that incorporate the hierarchical structure of *if-then* relation types lead to more accurate inference compared to models trained in isolation, as measured by both automatic and human evaluation.

Introduction

Given a snapshot observation of an event, people can easily anticipate and reason about unobserved causes and effects in relation to the observed event: what might have happened just before, what might happen next as a result, and how different events are chained through causes and effects. For instance, if we observe an event "X repels Y's attack" (Figure 1), we can immediately infer various plausible facts surrounding that event. In terms of the *plausible motivations* behind the event, X probably wants to protect herself. As for the *plausible pre-conditions* prior to the event, X may have been trained in self-defense to successfully fend off Y's attack. We can also infer the *plausible characteristics* of X; she might be strong, skilled, and brave. As a *result* of the event, X probably feels angry and might want to file a police report. Y on the other hand might feel scared of getting caught and want to run away.

The examples above illustrate how day-to-day commonsense reasoning can be operationalized through a densely connected collection of inferential knowledge. It is through this knowledge that we can watch a two hour movie and understand a story that spans over several months, as we can reason about a great number of events, causes, and effects,

Figure 1: A tiny subset of ATOMIC, an atlas of machine commonsense for everyday events, causes, and effects.

while observing only on a small fraction of them. It also enables us to develop Theories of Mind about others (Moore 2013). However, this ability, while common and trivial for humans, is what fundamentally lacks in today's AI systems. This is in part because the vast majority of AI systems are trained for task-specific datasets and objectives, which lead to models that are effective at finding task-specific correlation patterns but lack simple and explainable commonsense reasoning (Davis and Marcus 2015; Lake et al. 2017; Marcus 2018).

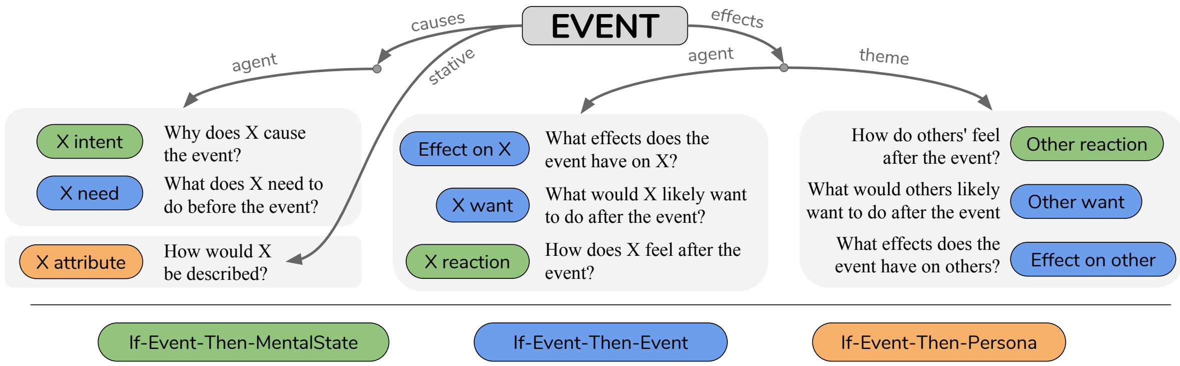

Figure 2: The taxonomy of *if-then* reasoning types. We consider nine *if-then* relations that have overlapping hierarchical structures as visualized above. One way to categorize the types is based on the type of content being predicted: (1) **If-Event-Then-Mental-State**, (2) **If-Event-Then-Event**, and (3) **If-Event-Then-Persona**. Another way is to categorize the types based on their causal relations: (1) **"causes"**, (2) **"effects"**, and (3) **"stative"**. Some of these categories can further divide depending on whether the reasoning focuses on the "agent" (X) or the "theme" (Other) of the event.

In this paper, we introduce ATOMIC,[1] an atlas of machine commonsense as a step toward addressing the rich spectrum of inferential knowledge that is crucial for automated commonsense reasoning. In contrast with previous efforts (Lenat 1995; Speer and Havasi 2012) that predominantly contain taxonomic or encyclopedic knowledge (Davis and Marcus 2015), ATOMIC focuses on inferential *if-then* knowledge. The goal of our study is to create a knowledge repository that meets three requirements: scale, coverage, and quality. Therefore, we focus on crowdsourcing experiments instead of extracting commonsense from corpora, because the latter is subject to the significant reporting bias in language (Gordon and Van Durme 2013) that can challenge both the coverage and quality of the extracted knowledge.

We propose a new taxonomy of *if-then* reasoning types as shown in Figure 2. One way to categorize the types is based on the content being predicted: (1) *If-Event-Then-Mental-State*, (2) *If-Event-Then-Event*, and (3) *If-Event-Then-Persona*. Another way to categorize is based on their causal relations: (1) "causes", (2) "effects", and (3) "stative". Using this taxonomy, we gather over 877K instances of inferential knowledge.

We then investigate neural network models that can acquire simple commonsense capabilities and reason about previously unseen events by embedding the rich inferential knowledge described in ATOMIC. Experimental results demonstrate that neural networks can abstract away commonsense inferential knowledge from ATOMIC such that given a previously unseen event, they can anticipate the likely causes and effects in rich natural language descriptions. In addition, we find that multitask models that can incorporate the hierarchical structure of if-then relation types lead to more accurate inference compared to models trained in isolation.

[1] an **AT**las **O**f **M**ach**I**ne **C**ommonsense, available to download or browse at `https://homes.cs.washington.edu/~msap/atomic/`

If-Then relation types

To enable better reasoning about events, we improve upon existing resources of commonsense knowledge by adding nine new causal and inferential dimensions. Shown in Figure 2, we define dimensions as denoting a particular type of *If-Then* knowledge, collected through crowdsourcing answers to questions about an event. Contrary to most previous work, ATOMIC also characterizes knowledge of events and their *implied* participants (e.g. "Alex calls for help" implies someone will answer the call), in addition to explicitly mentioned participants (e.g. "Alex calls Taylor for help").

Illustrated in Table 1, our nine dimensions span three types of *If-Then* relations, outlined below.

If-Event-Then-Mental-State We define three relations relating to the mental pre- and post-conditions of an event. Given an event (e.g., "X compliments Y"), we reason about (i) likely *intents* of the event (e.g., "X wants to be nice"), (ii) likely *(emotional) reactions* of the event's subject ("X feels good"), and (iii) likely *(emotional) reactions* of others ("Y feels flattered").

If-Event-Then-Event We also define six relations about probable pre- and post-conditions of events. Those relations describe events likely required to precede an event, as well as those likely to follow. For instance, people know that "X needs to put coffee in the filter" before "X makes Y's coffee". For post-conditions, we focus on both voluntary ("X adds cream and sugar") and involuntary ("X gets thanked by Y") possible next events. We also define voluntary and involuntary possible next events for (implied) participants.

If-Event-Then-Persona In addition to pre- and post-conditions, we also define a stative relation that describes how the subject of an event is described or perceived. For

Event	Type of relations	Inference examples	Inference dim.
"PersonX pays PersonY a compliment"	If-Event-Then-Mental-State	PersonX wanted to be nice	xIntent
		PersonX will feel good	xReact
		PersonY will feel flattered	oReact
	If-Event-Then-Event	PersonX will want to chat with PersonY	xWant
		PersonY will smile	oEffect
		PersonY will compliment PersonX back	oWant
	If-Event-Then-Persona	PersonX is flattering	xAttr
		PersonX is caring	xAttr
"PersonX makes PersonY's coffee"	If-Event-Then-Mental-State	PersonX wanted to be helpful	xIntent
		PersonY will be appreciative	oReact
		PersonY will be grateful	oReact
	If-Event-Then-Event	PersonX needs to put the coffee in the filter	xNeed
		PersonX gets thanked	xEffect
		PersonX adds cream and sugar	xWant
	If-Event-Then-Persona	PersonX is helpful	xAttr
		PersonX is deferential	xAttr
"PersonX calls the police"	If-Event-Then-Mental-State	PersonX wants to report a crime	xIntent
		Others feel worried	oReact
	If-Event-Then-Event	PersonX needs to dial 911	xNeed
		PersonX wants to explain everything to the police	xWant
		PersonX starts to panic	xEffect
		Others want to dispatch some officers	oWant
	If-Event-Then-Persona	PersonX is lawful	xAttr
		PersonX is responsible	xAttr

Table 1: Examples of **If-Event-Then-X** commonsense knowledge present in ATOMIC. For inference dimensions, "x" and "o" pertain to PersonX and others, respectively (e.g., "xAttr": attribute of PersonX, "oEffect": effect on others).

instance, when "X calls the police", X is seen as "lawful" or "responsible".

An Alternative Hierarchy The above relation types can be categorized via a different hierarchical structure as shown in Figure 2. In particular, they can be categorized based on their causal relations: (1) "causes", (2) "effects", and (3) "stative". Each of these categories can further divide depending on whether the reasoning focuses on the "agent" or the "theme" of the event. We omit cases where the combination is unlikely to lead to commonsense anticipation. For example, it is usually only the "agent" who causes the event, rather than the "theme", thus we don't consider that branching. We later consider this hierarchical structure of inferential relations for model effective neural network architectures that can learn to reason about a given event.

Data

To build ATOMIC, we create a crowdsourcing framework that allows for scalable, broad collection of *If-Then* knowledge for given events.

Compiling Base Events

As base events for our annotations, we extract 24K common event phrases from a variety of corpora. To ensure broad and diverse coverage, we compile common phrases from stories, books, Google Ngrams, and Wiktionary idioms (Mostafazadeh et al. 2016; Gordon and Swanson 2008; Goldberg and Orwant 2013). Following Rashkin et al. (2018), we define events as verb phrases with a verb predicate and its arguments ("drinks dark roast in the morning"). If a verb and its arguments do not co-occur frequently enough,[2] we replace the arguments with a blank placeholder ("drinks ___ in the morning"). In order to learn more general representations of events, we replace tokens referring to people with a `Person` variable (e.g. "PersonX buys PersonY coffee"). In future work, other types of variables could be added for other entity references (e.g. "PersonX moves to CityX").

For events with multiple people explicitly involved, we run a short annotation task to help resolve coreference chains within phrases. Disambiguating the participants is important, since it can drastically change the meaning of the event (e.g., "PersonX breaks PersonX's arm" vs. "PersonX breaks PersonY's arm" have very different implications). Three workers selected whether each "Person" mention in an event refers to PersonX, PersonY, or PersonZ, reaching an agreement of κ =0.4.

[2]We use frequency thresholds of 5 and 100 for stories and blogs, respectively, and limit ourselves to the top 10000 events in GoogleNgrams.

	count	# words
# triples: If-Event-Then-*	877,108	
-Event	521,334	
-MentalState	212,598	
-Persona	143,176	
# nodes: If-Event-Then-*	309,515	2.7
-Event	245,905	3.3
-MentalState	51,928	2.1
-Persona	11,495	1.0
base events	24,313	4.6
# nodes appearing >1	47,356	–

Table 2: Statistics of ATOMIC. Triples represent distinct `<event,relation,event>`

Crowdsourcing Framework

To ensure scalability, we implement a free-form text annotation setup which asks workers to simply write answers to questions about a specific event. We chose free-text over structured or categorical annotation for two reasons. First, categorical annotations with large labeling space has a substantial learning curve, which limits the annotation speed and thereby the coverage of our knowledge graph. Second, the categorical labels are likely to be fundamentally limiting to encode the vast space of commonsense knowledge and reasoning as depicted in Figure 1 and Table 1.

We create four tasks on Amazon Mechanical Turk (MTurk) for gathering commonsense annotations.[3] For each dimension, up to three workers are asked to provide as many as four likely annotations for an event, covering multiple possible situations (e.g., if "PersonX drinks coffee", then "PersonX needed to brew coffee" or "PersonX needed to buy coffee"; both are distinct but likely). Note that some events aren't caused by PersonX, and some don't affect other people, making annotations for certain dimensions not necessary (specifically, for xIntent, xNeed, oReact, oEffect, and oWant) for all events. For those dimensions, we first ask workers whether this specific inference dimension is relevant given an event.

ATOMIC Statistics

Table 2 lists descriptive statistics of our knowledge graph. Our resulting knowledge graph contains over 300K nodes, collected using 24K base events. Nodes in the graph are short phrases (2.7 tokens on average), ranging from 1 token for stative events (attributes) to 3.3 and 4.6 tokens for more active events. Our framework allows us to measure agreement on certain binary questions (e.g., whether an event affects others). On such questions, workers agreed moderately (pairwise percent agreement=75%, κ=0.37). Unlike denotational tasks where experts would only consider one label as correct, our annotations correspond to a distribution over *likely* inferences (de Marneffe, Manning, and Potts 2012). To confirm the correctness of our annotations, we run a small task asking MTurkers to determine whether an individual annotation is valid given a specific event and annotation. Table 4 shows that ATOMIC annotations are deemed valid 85% of the time for a random subset of events. Additionally, to further guarantee the quality of the free-form annotations, we manually filter out those by bad workers.

[3]The tasks were used to collect the following four sets of dimensions: (1) intent and reaction, (2) need and want, (3) effects, and (4) attributes.

Methods

Our goal is to investigate whether models can learn to perform *If-Then* commonsense inference given a previously unseen event. To this extent, we frame the problem as a conditional sequence generation problem: given an event phrase $\mathbf{e}$ and an inference dimension c, the model generates the target $\mathbf{t} = f_\theta(\mathbf{e}, \mathbf{c})$. Specifically, we explore various multitask encoder-decoder setups.

Encoder We represent the event phrase as a sequence of n word vectors $\mathbf{e} = \{e_0, e_1, \ldots, e_{n-1}\} \in \mathbb{R}^{n \times i_{enc}}$ where each word is an i_{enc}-dimensional vector. The event sequence is compressed into a hidden representation $\mathbf{h}$ through an encoding function $f_{enc} : \mathbb{R}^{i \times h_{enc}} \to \mathbb{R}^h$.

In this work, we use 300-dimensional static GloVe pretrained embeddings (Pennington, Socher, and Manning 2014) as our base word vectors. We augment these embeddings with 1024-dimensional ELMo pre-trained embeddings. ELMo (Peters et al. 2018) is a deep contextualized representation of words and it uses character-based representations, which allows robust representations of previously unseen events. The encoding function is a bidirectional GRU (Cho et al. 2014) of hidden size h_{enc}.

Decoder Each decoder is a unidirectional GRU of hidden size h_{dec}, with a hidden state initialized to $\mathbf{h}^{(0)}_{dec} = \mathbf{h}$. The target is represented by a sequence of vectors $\mathbf{t} = \{t_0, t_1, \ldots\}$, where each $t_i \in \mathbb{R}^h_{dec}$ based on a learned embedding. The decoder then maximizes $p(t_{i+1}|\mathbf{h}^{(i)}_{dec}, t_0, \ldots, t_i) = \text{softmax}(W_o \text{GRU}(\mathbf{h}^{(i)}_{dec}, t_i) + b_o)$

Single vs. Multitask Learning We experiment with various ways to combine the commonsense dimensions with multitask modelling. We design models that exploit the hierarchical structure of the commonsense dimensions (depicted in Figure 2), sharing encoders for dimensions that are related. Specifically, we explore the following models:

- EVENT2PERSONX/Y: we dissociate dimensions relating to the event's agent (PersonX) from those relating to the event's theme (others, or PersonY). This model has two encoders for six and three decoders, respectively.
- EVENT2(IN)VOLUNTARY: we further explore grouping dimensions together depending on whether they denote voluntary (e.g., xIntent, oWant) or involuntary (e.g., xReact, oEffect) events. This model connects two encoders to four "voluntary" and five "involuntary" decoders, respectively.

Dataset	Model	xIntent	xNeed	xAttr	xEffect	xWant	xReact	oEffect	oWant	oReact
DEV	9ENC9DEC	3.42	9.93	1.61	**7.26**	7.80	3.21	**4.81**	7.98	**3.89**
	EVENT2PRE/POST	3.35	9.83	–	7.23	8.36	3.26	4.74	8.50	3.74
	EVENT2(IN)VOLUNTARY	3.67	**11.62**	2.04	6.80	7.91	3.19	4.56	**9.05**	3.84
	EVENT2PERSONX/Y	**3.90**	11.04	**2.27**	6.52	**8.48**	**3.39**	4.43	7.92	3.55
TEST	9ENC9DEC	3.47	9.93	1.64	**7.53**	7.66	3.15	**5.02**	8.12	3.51
	EVENT2PRE/POST	3.48	**11.33**	-	7.08	8.63	3.28	4.53	8.12	3.58
	EVENT2(IN)VOLUNTARY	3.92	11.26	1.99	7.46	7.93	3.39	4.69	**8.85**	**3.72**
	EVENT2PERSONX/Y	**4.02**	11.31	**2.25**	6.55	**8.73**	**3.47**	4.52	7.71	3.62

Table 3: Average BLEU score for the top 10 predictions for each inference dimension - comparison of multitask models to uni-task model

Model	xIntent	xNeed	xAttr	xEffect	xWant	xReact	oEffect	oWant	oReact	average
9ENC9DEC	38.20	37.14	34.52	38.00	37.28	**53.02**	22.88	30.08	32.16	35.52
EVENT2PRE/POST	36.24	32.60	–	31.08	35.14	52.32	20.18	31.10	31.94	33.83
EVENT2(IN)VOLUNTARY	35.86	37.08	**34.78**	35.96	41.20	51.12	19.90	31.10	**32.80**	35.53
EVENT2PERSONX/Y	**39.82**	**43.90**	34.28	**39.56**	**46.56**	51.14	**25.34**	**34.12**	30.78	**38.39**
gold ATOMIC annotations	91.59	81.82	74.66	80.36	88.16	94.49	83.15	80.20	86.25	84.52

Table 4: Precision @ 10 (%) of generated inferences as selected by human judges for four models, averaged and broken down by dimension. We bold the best performing model for that dimension. EVENT2PERSONX/Y outperforms other models significantly ($p<0.001$).

- EVENT2PRE/POST: lastly, we split our dimensions based on whether they are related to causes (xNeed, xIntent) or effects (e.g., xWant, oEffect, xReact). In this model, there are two encoders and 8 decoders.[4]

As a single task baseline, we train nine separate encoder-decoders, one for each dimension (9ENC9DEC).

Training Details To test our zero-shot learning, we split seed events into training, validation, and test set (80/10/10). As is common in generation tasks, we minimize the cross entropy of the distribution over predicted targets compared to the gold distribution in our data.[5] During multitask training, we average the cross entropy of each task. Since multiple crowdworkers annotated each event, we define our training instances to be the combination of one worker's annotations. During experiments, we use the 300-d GloVe embeddings, yielding an encoder input size of $i_{enc} = 1324$ once concatenated with the 1,024-d ELMo embeddings. In the encoder, ELMo's character-level modelling allows for an unlimited vocabulary. For the decoder, we restrict the vocabulary to tokens appearing at least twice. We set the encoder and decoder hidden sizes to $h_{enc} = 100$ and $h_{dec} = 100$.

Results

We evaluate our various models for zero-shot reasoning about events; given a previously unseen event, a model generates the nine types of *if-then* inferences. We report performance using automatic metrics and a human evaluation of the generated inferences.

Automatic Metrics

We automatically evaluate the sequence generation for each model and each inference dimension using BLEU score. Specifically, we compute the average BLEU score ($n = 2$, `Smoothing1` (Chen and Cherry 2014)) between each sequence in the top 10 predictions and the corresponding set of MTurk annotations. Table 3 presents the results on both DEV and TEST datasets. The experiments show that models that exploit the hierarchical structure of the commonsense relations perform better than the model that uses separate parameters (9ENC9DEC). Importantly, BLEU is a crude measure of performance as it is based on the exact match of $n$grams. As shown in Figure 3, the generated samples depict varying word and phrase choices, thus we also perform human performance to complement automatic evaluations.

Human Evaluation

Since automatic evaluation of generated language is an open research question (Liu et al. 2016), we also assess our models' performance through human evaluation. We randomly select 100 events from the test set and use beam search to generate the 10 most likely inferences per dimension. We present five crowdworkers with the 10 decoded phrases, and ask workers to select all generations they think make sense. Table 4 shows each model's precision @ 10, computed as the average number of generations selected. As a comparison, we also assess the quality of the gold ATOMIC annota-

[4]We omit xAttr in this model, as it is trivially covered in the single task baseline.

[5]All our experiments were run using AllenNLP (Gardner et al. 2017).

Figure 3: Examples of generated inferences ( ) for three events from the test set, ordered from most likely (top) to least likely (bottom) according to the EVENT2PERSONX/Y model. Below the generations ( ), we show the crowd-sourced inferences.

tions for those events, asking five workers whether a particular annotation is valid for an event and a dimension. Shown in the last line of Table 4, workers deemed that 84.5% were valid, showcasing the quality of commonsense knowledge contained in ATOMIC.

As with automatic evaluation, models that leverage the commonsense hierarchy perform higher when rated by humans. Specifically, explicitly modelling whether an inference relates to the event's agent (X) or its theme (others) yields more sensible generations, as evidenced by the performance of EVENT2PERSONX/Y.

Qualitative Results

We present sample commonsense predictions in Figure 3. Given an event "PersonX starts lessons", our model can correctly infer that X probably *needs* to "practice" or "register for classes" or "find a tutor". Our model also correctly predicts that the likely effect of this event would be that X will "gain knowledge" or "learns a new skill". While our predicted inferences are closely related to the *Gold* descriptions, which are crowdsourced inferences, we can see that the words and phrases do not necessarily match exactly.

Comparison with ConceptNet

ConceptNet (Speer, Chin, and Havasi 2017) represents commonsense knowledge as a graph of *concepts* connected by *relations*. Concepts consist of words or phrases, while relations come from a fixed set of edge types.

While ConceptNet captures general commonsense knowledge – much of which is taxonomic in nature[6] – ATOMIC focuses on sequences of events and the social commonsense relating to them. This focus means that while events and dimensions in ATOMIC loosely correspond to concepts and relations from ConceptNet, individual dimensions, such as *intents*, can't be mapped cleanly onto any combination of ConceptNet's relations. The correspondence is neither one-to-one nor one-to-many. Still, in order to empirically investigate the differences between ConceptNet and ATOMIC, we used the following best-effort mappings between the dimensions and relations:

- **Intents**: MOTIVATEDBYGOAL, CAUSESDESIRE, HASSUBEVENT, HASFIRSTSUBEVENT
- **Reactions**: CAUSES, HASLASTSUBEVENT, HASSUBEVENT
- **Wants**: MOTIVATEDBYGOAL, HASSUBEVENT, HASFIRSTSUBEVENT, CAUSESDESIRE
- **Needs**: MOTIVATEDBYGOAL, ENTAILS, HASPREREQUISITE
- **Effects**: CAUSES, HASSUBEVENT, HASFIRSTSUBEVENT, HASLASTSUBEVENT
- **Attributes**: HASPROPERTY

We then computed the overlap of `<event1, dimension, event2>` triples in ATOMIC with the `<concept1, relation, concept2>` triples in ConceptNet. We found the overlap to only to be as high as 7% for *wants*, 6% for *effects*, 6% for *needs*, 5% for *intents*, 2% for reactions, and 0% for attributes. Moreover, only 25% of the events overall in ATOMIC are found in ConceptNet. Thus, ATOMIC offers a substantial amount of new inferential knowledge that has not been captured by existing resources.

Related work

Descriptive Knowledge from Crowdsourcing Knowledge acquisition and representation have been extensively

[6]While ConceptNet includes various inferential relations (e.g., entails, causes, motivated by), their instances amount to only about 1% of ConceptNet.

studied in prior research (Espinosa and Lieberman 2005; Speer and Havasi 2012; Lenat 1995). However, most prior efforts focused on taxonomic or encyclopedic knowledge (Davis and Marcus 2015), which, in terms of epistemology, corresponds to *knowledge of "what"*. Relatively less progress has been made on *knowledge of "how"* and *"why"*. For example, OpenCyc 4.0 is a large commonsense knowledge base consisting of 239,000 concepts and 2,039,000 facts (Lenat 1995) which have been noted to be mostly taxonomic (Davis and Marcus 2015).

Similarly, ConceptNet (Speer, Chin, and Havasi 2017) represents commonsense knowledge as a graph that connects words and phrases (*concepts*) with labeled edges (*relations*). For ConceptNet, inferential relations (e.g., "entails", "causes", "motivated by") amount to only about 1% of all triples in the graph. In contrast, ATOMIC is centered around events represented with natural language descriptions. While events and dimensions in ATOMIC loosely correspond to concepts and relations in ConceptNet, the two represent very different information and ultimately have relatively small overlap as discussed in the Results section.

The recent work of Rashkin et al. (2018) introduced a commonsense inference task about events and mental states: given an event described in natural language, the task is to generate the reaction and intent of actors involved in the event. ATOMIC is inspired by this work, but substantially scales up (i) the crowdsourcing procedure to nine dimensions per event, and (ii) the size of the knowledge graph – from 77K events in Event2Mind to 300K events in ATOMIC. Moreover, while the primary focus of (Rashkin et al. 2018) was inferential knowledge, its scope was limited to mental states.

Acquired Knowledge from Extraction and Induction More generally, the goal of moving beyond static commonsense knowledge to enable automated commonsense reasoning has inspired much research. Several projects have sought to extract commonsense inferential rules from naturally occurring resources such as large corpora (Schubert 2002), movie scripts (Tandon, de Melo, and Weikum 2017), and web how-tos (Chu, Tandon, and Weikum 2017). Such systems must inevitably deal with reporting bias (Gordon and Van Durme 2013), or the fact that the frequency and selection of phenomena represented in natural language systematically differs from what occurs in the real world. Other approaches have sought to induce commonsense rules from large knowledge bases (Galárraga et al. 2013; Yang et al. 2015). While these approaches have also had success, the choice of schema and information represented in current knowledge bases limits the scope of propositions such systems can learn.

Scripts and Narrative Reasoning Other work has focused more specifically on representing and reasoning about sequences of events, similarly to ATOMIC. Early work on event sequences studied *scripts*, a kind of structured representation for prototypical sequences of events (Schank and Abelson 1977). More recently, *narrative event chains* have been proposed as a similar formalism for prototypical sequences of events that may be learned from raw text (Chambers and Jurafsky 2008). This work additionally proposed the *Narrative Cloze Test* as a benchmark for story understanding. In contrast to narrative event chains, the *ROC Stories Corpus* crowdsources event sequences represented as natural language stories rather than using a specific formalism (Mostafazadeh et al. 2016). Additionally, the *Story Cloze Test* adapts these stories into a new benchmark by requiring systems to choose between the true and a false ending to the story. Our work interpolates between these two approaches by representing events in natural language while structuring the relationships between events into the edges of a graph. The *Choice of Plausible Alternatives* (COPA) task offers a similar benchmark for commonsense understanding of events and their relationships (Roemmele, Bejan, and Gordon 2011). In COPA, a system is presented a premise and two alternatives that might have a causal relationship with the premise. While COPA, like ATOMIC, represents events as free-form text with structured relationships, it covers only a limited number of relations (cause and effect) and is smaller scale having only 1000 instances.

Conclusion

We present ATOMIC – an atlas of everyday commonsense inferential knowledge about events described in natural language and associated with typed *if-then* relations. ATOMIC consists of over 300K events associated with nine relation types – making it the largest knowledge graph of its kind. Our crowdsourcing framework gathers annotations in the form of free-form textual responses to simple questions which enables large-scale high quality collection of commonsense about events. We also present neural network models that can learn to reason about previously unseen events to generate their likely causes and effects in natural language.

References

Chambers, N., and Jurafsky, D. 2008. Unsupervised learning of narrative event chains. In *ACL*.

Chen, B., and Cherry, C. 2014. A systematic comparison of smoothing techniques for sentence-level bleu. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 362–367.

Cho, K.; van Merrienboer, B.; Bahdanau, D.; and Bengio, Y. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *SSST@EMNLP*.

Chu, C. X.; Tandon, N.; and Weikum, G. 2017. Distilling task knowledge from how-to communities. In *WWW*.

Davis, E., and Marcus, G. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM* 58:92–103.

de Marneffe, M.-C.; Manning, C. D.; and Potts, C. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Comput. Linguist.* 38(2):301–333.

Espinosa, J. H., and Lieberman, H. 2005. Eventnet: Inferring temporal relations between commonsense events. In *MICAI*.

Galárraga, L.; Teflioudi, C.; Hose, K.; and Suchanek, F. M. 2013. Amie: association rule mining under incomplete evidence in ontological knowledge bases. In *WWW*.

Gardner, M.; Grus, J.; Neumann, M.; Tafjord, O.; Dasigi, P.; Liu, N. F.; Peters, M.; Schmitz, M.; and Zettlemoyer, L. S. 2017. Allennlp: A deep semantic natural language processing platform.

Goldberg, Y., and Orwant, J. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *SEM2013*.

Gordon, A. S., and Swanson, R. 2008. StoryUpgrade: finding stories in internet weblogs. In *ICWSM*.

Gordon, J., and Van Durme, B. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, 25–30. New York, NY, USA: ACM.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *The Behavioral and brain sciences* 40:e253.

Lenat, D. B. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM* 38(11):33–38.

Liu, C.-W.; Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *ACL*.

Marcus, G. 2018. Deep learning: A critical appraisal. *CoRR* abs/1801.00631.

Moore, C. 2013. *The development of commonsense psychology*. Psychology Press.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Rashkin, H.; Sap, M.; Allaway, E.; Smith, N. A.; and Choi, Y. 2018. Event2mind: Commonsense inference on events, intents, and reactions. In *ACL*.

Roemmele, M.; Bejan, C. A.; and Gordon, A. S. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Schank, R., and Abelson, R. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. The Artificial Intelligence Series. Lawrence Erlbaum Associates.

Schubert, L. 2002. Can we derive general world knowledge from texts? In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, 94–97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Speer, R., and Havasi, C. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 4444–4451.

Tandon, N.; de Melo, G.; and Weikum, G. 2017. Webchild 2.0 : Fine-grained commonsense knowledge distillation. In *ACL*.

Yang, B.; Yih, S. W.-t.; He, X.; Gao, J.; and Deng, L. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*.