# Event2Mind: Commonsense Inference on Events, Intents, and Reactions

**Hannah Rashkin**[†*] **Maarten Sap**[†*] **Emily Allaway**[†] **Noah A. Smith**[†] **Yejin Choi**[†‡]
[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence
{hrashkin,msap,eallaway,nasmith,yejin}@cs.washington.edu

## Abstract

We investigate a new commonsense inference task: given an event described in a short free-form text ("X drinks coffee in the morning"), a system reasons about the likely intents ("X wants to stay awake") and reactions ("X feels alert") of the event's participants. To support this study, we construct a new crowdsourced corpus of 25,000 event phrases covering a diverse range of everyday events and situations. We report baseline performance on this task, demonstrating that neural encoder-decoder models can successfully compose embedding representations of previously unseen events and reason about the likely intents and reactions of the event participants. In addition, we demonstrate how commonsense inference on people's intents and reactions can help unveil the implicit gender inequality prevalent in modern movie scripts.

## 1 Introduction

Understanding a narrative requires commonsense reasoning about the mental states of people in relation to events. For example, if "Alex is dragging his feet at work", pragmatic implications about Alex's *intent* are that "Alex wants to avoid doing things" (Figure 1). We can also infer that Alex's *emotional reaction* might be feeling "lazy" or "bored". Furthermore, while not explicitly mentioned, we can infer that people other than Alex are affected by the situation, and these people are likely to feel "frustrated" or "impatient".

This type of pragmatic inference can potentially be useful for a wide range of NLP applications

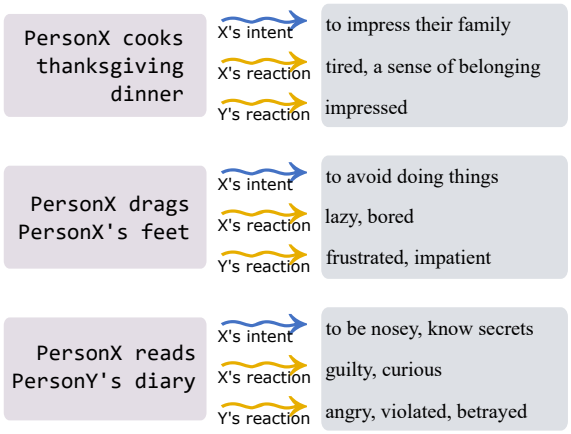


Figure 1: Examples of commonsense inference on mental states of event participants. In the third example event, common sense tells us that Y is likely to feel betrayed as a result of X reading their diary.

that require accurate anticipation of people's intents and emotional reactions, even when they are not explicitly mentioned. For example, an ideal dialogue system should react in empathetic ways by reasoning about the human user's mental state based on the events the user has experienced, without the user explicitly stating how they are feeling. Similarly, advertisement systems on social media should be able to reason about the emotional reactions of people after events such as mass shootings and remove ads for guns which might increase social distress (Goel and Isaac, 2016). Also, pragmatic inference is a necessary step toward automatic narrative understanding and generation (Tomai and Forbus, 2010; Ding and Riloff, 2016; Ding et al., 2017). However, this type of social commonsense reasoning goes far beyond the widely studied entailment tasks (Bowman et al., 2015; Dagan et al., 2006) and thus falls outside the scope of existing benchmarks.

In this paper, we introduce a new task, corpus,

[*]These two authors contributed equally.

| PersonX's Intent | Event Phrase | PersonX's Reaction | Others' Reactions |
|---|---|---|---|
| to express anger<br>to vent their frustration<br>to get PersonY's full attention | **PersonX starts to yell at PersonY** | mad<br>frustrated<br>annoyed | shocked<br>humiliated<br>mad at PersonX |
| to communicate something without being rude<br>to let the other person think for themselves<br>to be subtle | **PersonX drops a hint** | sly<br>secretive<br>frustrated | oblivious<br>surprised<br>grateful |
| to catch the criminal<br>to be civilized<br>justice | **PersonX reports ___ to the police** | anxious<br>worried<br>nervous | sad<br>angry<br>regret |
| to wake up<br>to feel more energized | **PersonX drinks a cup of coffee** | alert<br>awake<br>refreshed | NONE |
| to be feared<br>to be taken seriously<br>to exact revenge | **PersonX carries out PersonX's threat** | angry<br>dangerous<br>satisfied | sad<br>afraid<br>angry |
| NONE | **It starts snowing** | NONE | calm<br>peaceful<br>cold |

Table 1: Example annotations of intent and reactions for 6 event phrases. Each annotator could fill in up to three free-responses for each mental state.

and model, supporting commonsense inference on events with a specific focus on modeling stereotypical intents and reactions of people, described in short free-form text. Our study is in a similar spirit to recent efforts of Ding and Riloff (2016) and Zhang et al. (2017), in that we aim to model aspects of commonsense inference via natural language descriptions. Our new contributions are: (1) a new corpus that supports commonsense inference about people's intents and reactions over a diverse range of everyday events and situations, (2) inference about even those people who are not directly mentioned by the event phrase, and (3) a task formulation that aims to *generate* the textual descriptions of intents and reactions, instead of classifying their polarities or classifying the inference relations between two given textual descriptions.

Our work establishes baseline performance on this new task, demonstrating that, given the phrase-level inference dataset, neural encoder-decoder models can successfully compose phrasal embeddings for previously unseen events and reason about the mental states of their participants.

Furthermore, in order to showcase the practical implications of commonsense inference on events and people's mental states, we apply our model to modern movie scripts, which provide a new insight into the gender bias in modern films beyond what previous studies have offered (England et al., 2011; Agarwal et al., 2015; Ramakrishna et al., 2017; Sap et al., 2017). The resulting corpus includes around 25,000 event phrases, which combine automatically extracted phrases from stories and blogs with all idiomatic verb phrases listed in the Wiktionary. Our corpus is publicly available.[1]

## 2 Dataset

One goal of our investigation is to probe whether it is feasible to build computational models that can perform limited, but well-scoped commonsense inference on short free-form text, which we refer to as *event phrases*. While there has been much prior research on phrase-level paraphrases (Pavlick et al., 2015) and phrase-level entailment (Dagan et al., 2006), relatively little prior work focused on phrase-level inference that requires prag-

[1] https://tinyurl.com/event2mind

matic or commonsense interpretation. We scope our study to two distinct types of inference: given a phrase that describes an event, we want to reason about the likely intents and emotional reactions of people who caused or affected by the event. This complements prior work on more general commonsense inference (Speer and Havasi, 2012; Li et al., 2016; Zhang et al., 2017), by focusing on the causal relations between events and people's mental states, which are not well covered by most existing resources.

We collect a wide range of phrasal event descriptions from stories, blogs, and Wiktionary idioms. Compared to prior work on phrasal embeddings (Wieting et al., 2015; Pavlick et al., 2015), our work generalizes the phrases by introducing (typed) variables. In particular, we replace words that correspond to entity mentions or pronouns with typed variables such as `PersonX` or `PersonY`, as shown in examples in Table 1. More formally, the phrases we extract are a combination of a verb predicate with partially instantiated arguments. We keep specific arguments together with the predicate, if they appear frequently enough (e.g., `PersonX eats pasta for dinner`). Otherwise, the arguments are replaced with an untyped blank (e.g., `PersonX eats ___ for dinner`). In our work, only person mentions are replaced with typed variables, leaving other types to future research.

**Inference types** The first type of pragmatic inference is about *intent*. We define intent as an explanation of why the agent causes a volitional event to occur (or "none" if the event phrase was unintentional). The intent can be considered a mental pre-condition of an action or an event. For example, if the event phrase is `PersonX takes a stab at ___`, the annotated intent might be that "PersonX wants to solve a problem".

The second type of pragmatic inference is about *emotional reaction*. We define reaction as an explanation of how the mental states of the agent and other people involved in the event would change as a result. The reaction can be considered a mental post-condition of an action or an event. For example, if the event phrase is that `PersonX gives PersonY ___ as a gift`, PersonX might "feel good about themselves" as a result, and PersonY might "feel grateful" or "feel thankful".

| Source | # Unique Events | # Unique Verbs | Average $\kappa$ |
|---|---|---|---|
| ROC Story | 13,627 | 639 | 0.57 |
| G. N-grams | 7,066 | 789 | 0.39 |
| Spinn3r | 2,130 | 388 | 0.41 |
| Idioms | 1,916 | 442 | 0.42 |
| **Total** | **24,716** | **1,333** | **0.45** |

Table 2: Data and annotation agreement statistics for our new phrasal inference corpus. Each event is annotated by three crowdworkers.

### 2.1 Event Extraction

We extract phrasal events from three different corpora for broad coverage: the ROC Story training set (Mostafazadeh et al., 2016), the Google Syntactic N-grams (Goldberg and Orwant, 2013), and the Spinn3r corpus (Gordon and Swanson, 2008). We derive events from the set of verb phrases in our corpora, based on syntactic parses (Klein and Manning, 2003). We then replace the predicate subject and other entities with the typed variables (e.g., `PersonX`, `PersonY`), and selectively substitute verb arguments with blanks (___). We use frequency thresholds to select events to annotate (for details, see Appendix A.1). Additionally, we supplement the list of events with all 2,000 verb idioms found in Wiktionary, in order to cover events that are less compositional.[2] Our final annotation corpus contains nearly 25,000 event phrases, spanning over 1,300 unique verb predicates (Table 2).

### 2.2 Crowdsourcing

We design an Amazon Mechanical Turk task to annotate the mental pre- and post-conditions of event phrases. A snippet of our MTurk HIT design is shown in Figure 2. For each phrase, we ask three annotators whether the agent of the event, PersonX, intentionally causes the event, and if so, to provide up to three possible textual descriptions of their intents. We then ask annotators to provide up to three possible reactions that PersonX might experience as a result. We also ask annotators to provide up to three possible reactions of *other people*, when applicable. These other people can be either explicitly mentioned (e.g., "PersonY" in `PersonX punches PersonY's lights out`), or only implied

[2] We compiled the list of idiomatic verb phrases by cross-referencing between Wiktionary's English idioms category and the Wiktionary English verbs categories.

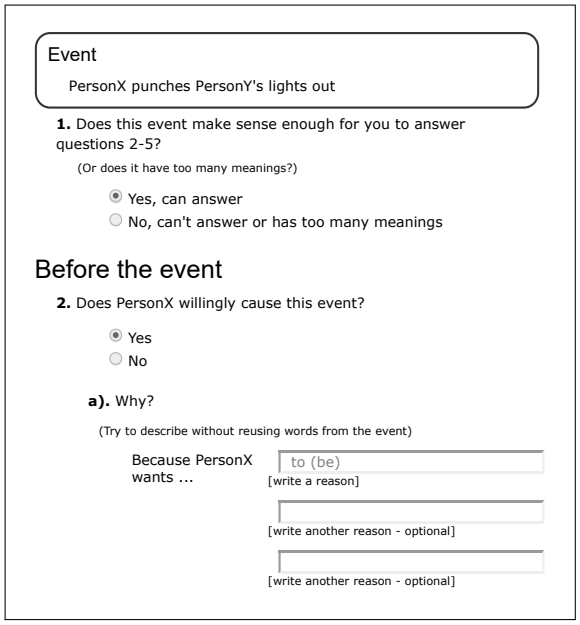


Figure 2: *Intent* portion of our annotation task. We allow annotators to label events as invalid if the phrase is unintelligible. The full annotation setup is shown in Figure 8 in the appendix.

(e.g., given the event description `PersonX yells at the classroom`, we can infer that other people such as "students" in the classroom may be affected by the act of PersonX). For quality control, we periodically removed workers with high disagreement rates, at our discretion.

**Coreference among `Person` variables** With the typed `Person` variable setup, events involving multiple people can have multiple meanings depending on coreference interpretation (e.g., `PersonX eats PersonY's lunch` has very different mental state implications from `PersonX eats PersonX's lunch`). To prune the set of events that will be annotated for intent and reaction, we ran a preliminary annotation to filter out candidate events that have implausible coreferences. In this preliminary task, annotators were shown a combinatorial list of coreferences for an event (e.g., `PersonX punches PersonX's lights out`, `PersonX punches PersonY's lights out`) and were asked to select only the plausible ones (e.g., `PersonX punches PersonY's lights out`). Each set of coreferences was annotated by 3 workers, yielding an overall agreement of $\kappa$ =0.4. This annotation excluded 8,406 events with implausible coreference from our set (out of 17,806 events).

### 2.3 Mental State Descriptions

Our dataset contains nearly 25,000 event phrases, with annotators rating 91% of our extracted events as "valid" (i.e., the event makes sense). Of those events, annotations for the multiple choice portions of the task (whether or not there exists intent/reaction) agree moderately, with an average Cohen's $\kappa = 0.45$ (Table 2). The individual $\kappa$ scores generally indicate that turkers disagree half as often as if they were randomly selecting answers.

Importantly, this level of agreement is acceptable in our task formulation for two reasons. First, unlike linguistic annotations on syntax or semantics where experts in the corresponding theory would generally agree on a single correct label, pragmatic interpretations may better be defined as distributions over multiple correct labels (e.g., after `PersonX takes a test`, PersonX might feel relieved and/or stressed; de Marneffe et al., 2012). Second, because we formulate our task as a conditional language modeling problem, where a distribution over the textual descriptions of intents and reactions is conditioned on the event description, this variation in the labels is only as expected.

A majority of our events are annotated as willingly caused by the agent (86%, Cohen's $\kappa = 0.48$), and 26% involve other people ($\kappa = 0.41$). Most event patterns in our data are fully instantiated, with only 22% containing blanks (___). In our corpus, the intent annotations are slightly longer (3.4 words on average) than the reaction annotations (1.5 words).

## 3 Models

Given an event phrase, our models aim to generate three entity-specific pragmatic inferences: PersonX's intent, PersonX's reaction, and others' reactions. The general outline of our model architecture is illustrated in Figure 3.

The input to our model is an event pattern described through free-form text with typed variables such as `PersonX gives PersonY ___ as a gift`. For notation purposes, we describe each event pattern $E$ as a sequence of word embeddings $\langle e_1, e_2, \ldots, e_n \rangle \in \mathbb{R}^{n \times D}$. This input is encoded as a vector $h_E \in \mathbb{R}^H$ that will be used for predicting output. The output of the model is its hypotheses about PersonX's intent, PersonX's reaction, and others' reactions ($v_i$,$v_x$, and $v_o$, respectively). We experiment with representing the

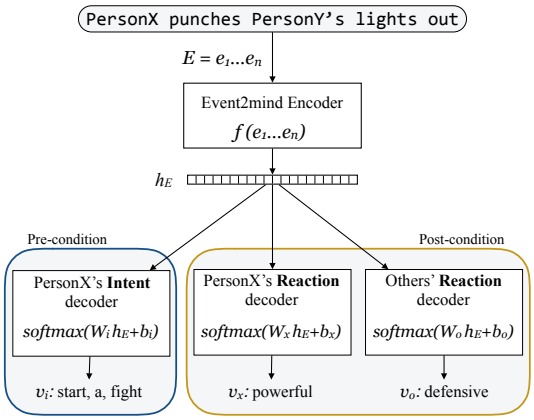


Figure 3: Overview of the model architecture. From an encoded event, our model predicts intents and reactions in a multitask setting.

output in two decoding set-ups: three vectors interpretable as discrete distributions over words and phrases (n-gram reranking) or three sequences of words (sequence decoding).

**Encoding events** The input event phrase $E$ is compressed into an $H$-dimensional embedding $h_E$ via an encoding function $f : \mathbb{R}^{n \times D} \rightarrow \mathbb{R}^H$:

$$h_E = f(e_1, \ldots, e_n)$$

We experiment with several ways for defining $f$, inspired by standard techniques in sentence and phrase classification (Kim, 2014). First, we experiment with max-pooling and mean-pooling over the word vectors $\{e_i\}_{i=1}^n$. We also consider a convolutional neural network (ConvNet; LeCun et al., 1998) taking the last layer of the network as the encoded version of the event. Lastly, we encode the event phrase with a bi-directional RNN (specifically, a GRU; Cho et al., 2014), concatenating the final hidden states of the forward and backward cells as the encoding: $h_E = [\overrightarrow{h_n}; \overleftarrow{h_1}]$. For hyperparameters and other details, we refer the reader to Appendix B.

Though the event sequences are typically rather short (4.6 tokens on average), our model still benefits from the ConvNet and BiRNN's ability to compose words.

**Pragmatic inference decoding** We use three decoding modules that take the event phrase embedding $h_E$ and output distributions of possible PersonX's intent ($v_i$), PersonX's reactions ($v_x$), and others' reactions ($v_o$). We experiment with two different decoder set-ups.

First, we experiment with *n-gram re-ranking*, considering the $|V|$ most frequent $\{1, 2, 3\}$-grams in our annotations. Each decoder projects the event phrase embedding $h_E$ into a $|V|$-dimensional vector, which is then passed through a softmax function. For instance, the distribution over descriptions of PersonX's intent is given by:

$$v_i = \text{softmax}(W_i h_E + b_i)$$

Second, we experiment with *sequence generation*, using RNN decoders to generate the textual description. The event phrase embedding $h_E$ is set as the initial state $h_{dec}$ of three decoder RNNs (using GRU cells), which then output the intent/reactions one word at a time (using beam-search at test time). For example, an event's intent sequence $(v_i = v_i^{(0)} v_i^{(1)} \ldots)$ is computed as follows:

$$v_i^{(t+1)} = \text{softmax}(W_i \text{ RNN}(v_i^{(t)}, h_{i,dec}^{(t)}) + b_i)$$

**Training objective** We minimize the cross-entropy between the predicted distribution over words and phrases, against the one actually observed in our dataset. Further, we employ multitask learning, simultaneously minimizing the loss for all three decoders at each iteration.

**Training details** We fix our input embeddings, using 300-dimensional skip-gram word embeddings trained on Google News (Mikolov et al., 2013). For decoding, we consider a vocabulary of size $|V| = 14{,}034$ in the n-gram re-ranking setup. For the sequence decoding setup, we only consider the unigrams in $V$, yielding an output space of 7,110 at each time step.

We randomly divided our set of 24,716 unique events (57,094 annotations) into a training/dev./test set using an 80/10/10% split. Some annotations have multiple responses (i.e., a crowdworker gave multiple possible intents and reactions), in which case we take each of the combinations of their responses as a separate training example.

## 4 Empirical Results

Table 3 summarizes the performance of different encoding models on the dev and test set in terms of cross-entropy and recall at 10 predicted intents and reactions. As expected, we see a moderate improvement in recall and cross-entropy when using the more compositional encoder models (ConvNet and BiRNN; both n-gram and sequence de-

| Encoding Function | Decoder | Development | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Average Cross-Ent | Recall @10 (%) Intent | XReact | OReact | Average Cross-Ent | Recall @10 (%) Intent | XReact | OReact |
| max-pool | n-gram | 5.75 | 31 | 35 | 68 | 5.14 | 31 | 37 | 67 |
| mean-pool | n-gram | 4.82 | 35 | 39 | 69 | 4.94 | 34 | 40 | 68 |
| ConvNet | n-gram | 4.85 | 36 | 42 | 69 | 4.81 | 37 | 44 | 69 |
| BiRNN 300d | n-gram | 4.78 | 36 | 42 | 68 | 4.74 | 36 | 43 | 69 |
| BiRNN 100d | n-gram | 4.76 | 36 | 41 | 68 | 4.73 | 37 | 43 | 68 |
| mean-pool | sequence | 4.59 | 39 | 36 | 67 | 4.54 | 40 | 38 | 66 |
| ConvNet | sequence | 4.44 | 42 | 39 | 68 | 4.40 | 43 | 40 | 67 |
| BiRNN 100d | sequence | 4.25 | 39 | 38 | 67 | 4.22 | 40 | 40 | 67 |

Table 3: Average cross-entropy (lower is better) and recall @10 (percentage of times the gold falls within the top 10 decoded; higher is better) on development and test sets for different modeling variations. We show recall values for PersonX's intent, PersonX's reaction and others' reaction (denoted as "Intent", "XReact", and "OReact"). Note that because of two different decoding setups, cross-entropy between n-gram and sequence decoding are not directly comparable.

coding setups). Additionally, BiRNN models outperform ConvNets on cross-entropy in both decoding setups. Looking at the recall split across intent vs. reaction labels ("Intent", "XReact" and "OReact" columns), we see that much of the improvement in using these two models is within the prediction of PersonX's intents. Note that recall for "OReact" is much higher, since a majority of events do not involve other people.

**Human evaluation** To further assess the quality of our models, we randomly select 100 events from our test set and ask crowd-workers to rate generated intents and reactions. We present 5 workers with an event's top 10 most likely intents and reactions according to our model and ask them to select all those that make sense to them. We evaluate each model's precision @10 by computing the average number of generated responses that make sense to annotators.

Figure 4 summarizes the results of this evaluation. In most cases, the performance is higher for the sequential decoder than the corresponding n-gram decoder. The biggest gain from using sequence decoders is in intent prediction, possibly because intent explanations are more likely to be longer. The BiRNN and ConvNet encoders consistently have higher precision than the mean-pooling with the BiRNN-seq setup slightly outperforming other models. Unless otherwise specified, this is the model we employ in further sections.

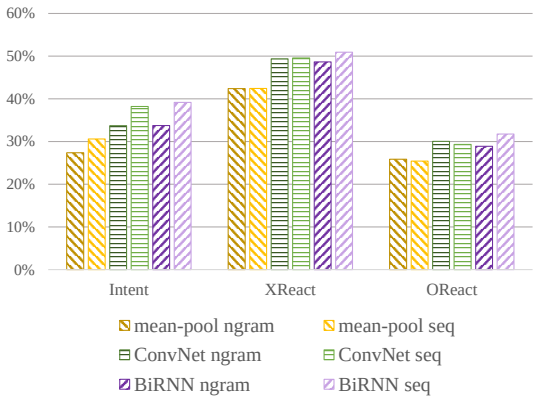


Figure 4: Average precision @10 of each model's top ten responses in the human evaluation. We show results for various encoder functions (mean-pool, ConvNet, BiRNN-100d) combined with two decoding setups (n-gram re-ranking, sequence generation).

**Error analyses** We test whether certain types of events are easier for predicting commonsense inference. In Figure 6, we show the difference in cross-entropy of the BiRNN 100d model on predicting different portions of the development set including: `Blank` events (events containing non-instantiated arguments), `2+ People` events (events containing multiple different Person variables), and `Idiom` events (events coming from the Wiktionary idiom list). Our results show that, while intent prediction performance remains sim-

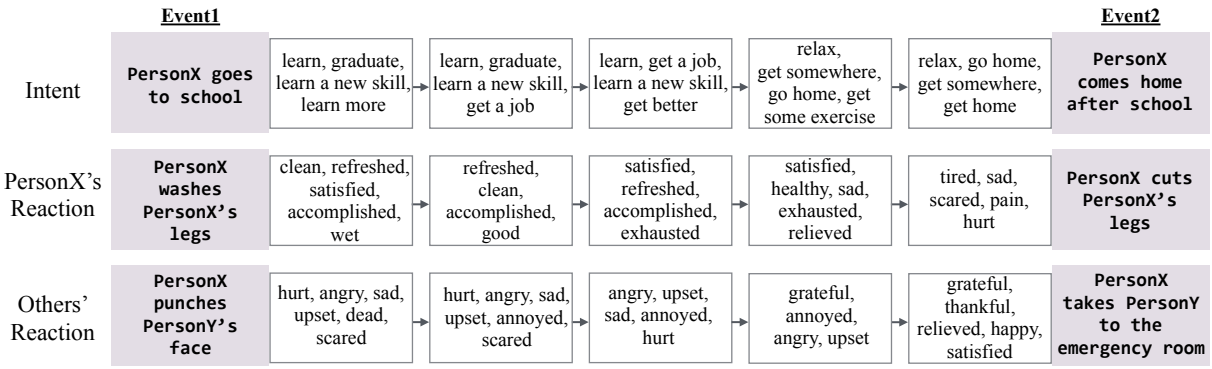


Figure 5: Sample predictions from homotopic embeddings (gradual interpolation between Event1 and Event2), selected from the top 10 beam elements decoded in the sequence generation setup. Examples highlight differences captured when ideas are similar (*going to* and *coming from* school), when only a single word differs (*washes* versus *cuts*), and when two events are unrelated.

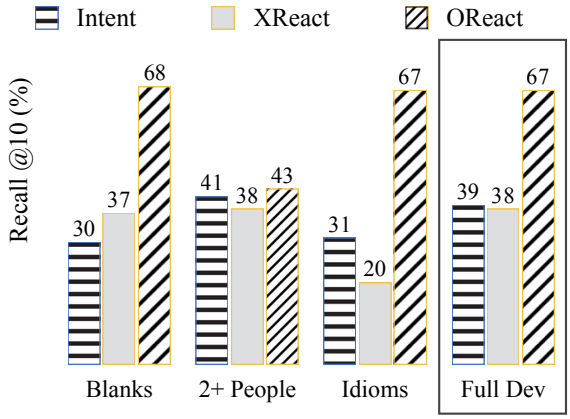


Figure 6: Recall @ 10 (%) on different subsets of the development set for intents, PersonX's reactions, and other people's reactions, using the BiRNN 100d model. "Full dev" represents the recall on the entire development dataset.

ilar for all three sets of events, it is 10% behind intent prediction on the full development set. Additionally, predicting other people's reactions is more difficult for the model when other people are explicitly mentioned. Unsurprisingly, idioms are particularly difficult for commonsense inference, perhaps due to the difficulty in composing meaning over nonliteral or noncompositional event descriptions.

To further evaluate the geometry of the embedding space, we analyze interpolations between pairs of event phrases (from outside the train set), similar to the homotopic analysis of Bowman et al. (2016). For a handful of event pairs, we decode intents, reactions for PersonX, and reactions for other people from points sampled at equal intervals on the interpolated line between two event phrases. We show examples in Figure 5. The embedding space distinguishes changes from generally positive to generally negative words and is also able to capture small differences between event phrases (such as "washes" versus "cuts").

## 5 Analyzing Bias via Event2Mind Inference

Through Event2Mind inference, we can attempt to bring to the surface what is implied about people's behavior and mental states. We employ this inference to analyze implicit bias in modern films. As shown in Figure 7, our model is able to analyze character portrayal beyond what is explicit in text, by performing pragmatic inference on character actions to explain aspects of a character's mental state. In this section, we use our model's inference to shed light on gender differences in intents behind and reactions to characters' actions.

### 5.1 Processing of Movie Scripts

For our portrayal analyses, we use scene descriptions from 772 movie scripts released by Gorinski and Lapata (2015), assigned to over 21,000 characters as done by Sap et al. (2017). We extract events from the scene descriptions, and generate their 10 most probable intent and reaction sequences using our BiRNN sequence model (as in Figure 7).

We then categorize generated intents and reactions into groups based on LIWC category scores of the generated output (Tausczik and Pennebaker, 2016).[3] The intent and reaction categories are then

[3]We only consider content word categories: 'Core Drives

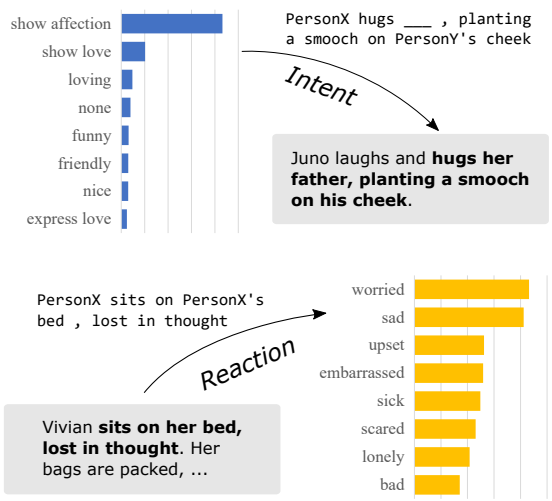


Figure 7: Two scene description snippets from *Juno* (2007, top) and *Pretty Woman* (1990, bottom), augmented with Event2mind inferences on the characters' intents and reactions. E.g., our model infers that the event `PersonX sits on PersonX's bed, lost in thought` implies that the agent, Vivian, is sad or worried.

aggregated for each character, and standardized (zero-mean and unit variance).

We compute correlations with gender for each category of intent or reaction using a logistic regression model, testing significance while using Holm's correction for multiple comparisons (Holm, 1979).[4] To account for the gender skew in scene presence (29.4% of scenes have women), we statistically control for the total number of words in a character's scene descriptions. Note that the original event phrases are all gender agnostic, as their participants have been replaced by variables (e.g., `PersonX`). We also find that the types of gender biases uncovered remain similar when we run these analyses on the human annotations or the generated words and phrases from the BiRNN with n-gram re-ranking decoding setup.

and Needs', 'Personal Concerns', 'Biological Processes', 'Cognitive Processes', 'Social Words', 'Affect Words', 'Perceptual Processes'. We refer the reader to Tausczik and Pennebaker (2016) or `http://liwc.wpengine.com/compare-dictionaries/` for a complete list of category descriptions.

[4]Given the data limitation, we represent gender as a binary, but acknowledge that gender is a more complex social construct.

### 5.2 Revealing Implicit Bias via Explicit Intents and Reactions

| |
|---|
| **Female: intents** |
| AFFILIATION, FRIEND, FAMILY |
| BODY, SEXUAL, INGEST |
| SEE, INSIGHT, DISCREP |
| **Male: intents** |
| DEATH, HEALTH, ANGER, NEGEMO |
| RISK, POWER, ACHIEVE, REWARD, WORK |
| CAUSE, TENTATIVE$^{\ddagger}$ |
| **Female: reactions** |
| POSEMO, AFFILIATION, FRIEND, REWARD |
| INGEST, SEXUAL$^{\ddagger}$, BODY$^{\ddagger}$ |
| **Male: reactions** |
| WORK, ACHIEVE, POWER, HEALTH$^{\dagger}$ |
| **Female: others' reactions** |
| POSEMO, AFFILIATION, FRIEND |
| INGEST, SEE, INSIGHT |
| **Male: others' reactions** |
| ACHIEVE, RISK$^{\dagger}$ |
| SAD, NEGEMO$^{\ddagger}$, ANGER$^{\dagger}$ |

Table 4: Select LIWC categories correlated with gender. All results are significant when corrected for multiple comparisons at $p < 0.001$, except $^{\dagger}p < 0.05$ and $^{\ddagger}p < 0.01$.

Our Event2Mind inferences automate portrayal analyses that previously required manual annotations (Behm-Morawitz and Mastro, 2008; Prentice and Carranza, 2002; England et al., 2011). Shown in Table 4, our results indicate a gender bias in the behavior ascribed to characters, consistent with psychology and gender studies literature (Collins, 2011). Specifically, events with female semantic agents are intended to be helpful to other people (intents involving FRIEND, FAMILY, and AFFILIATION), particularly relating to eating and making food for themselves and others (INGEST, BODY). Events with male agents on the other hand are motivated by and resulting in achievements (ACHIEVE, MONEY, REWARDS, POWER).

Women's looks and sexuality are also emphasized, as their actions' intents and reactions are sexual, seen, or felt (SEXUAL, SEE, PERCEPT). Men's actions, on the other hand, are motivated by violence or fighting (DEATH, ANGER, RISK), with strong negative reactions (SAD, ANGER, NEGATIVE EMOTION).

Our approach decodes nuanced implications

into more explicit statements, helping to identify and explain gender bias that is prevalent in modern literature and media. Specifically, our results indicate that modern movies have the bias to portray female characters as having pro-social attitudes, whereas male characters are portrayed as being competitive or pro-achievement. This is consistent with gender stereotypes that have been studied in movies in both NLP and psychology literature (Agarwal et al., 2015; Madaan et al., 2017; Prentice and Carranza, 2002; England et al., 2011).

## 6 Related Work

Prior work has sought formal frameworks for inferring roles and other attributes in relation to events (Baker et al., 1998; Das et al., 2014; Schuler et al., 2009; Hartshorne et al., 2013, *inter alia*), implicitly connoted by events (Reisinger et al., 2015; White et al., 2016; Greene, 2007; Rashkin et al., 2016), or sentiment polarities of events (Ding and Riloff, 2016; Choi and Wiebe, 2014; Russo et al., 2015; Ding and Riloff, 2018). In addition, recent work has studied the patterns which evoke certain polarities (Reed et al., 2017), the desires which make events affective (Ding et al., 2017), the emotions caused by events (Vu et al., 2014), or, conversely, identifying events or reasoning behind particular emotions (Gui et al., 2017). Compared to this prior literature, our work uniquely learns to model intents and reactions over a diverse set of events, includes inference over event participants not explicitly mentioned in text, and formulates the task as predicting the textual descriptions of the implied commonsense instead of classifying various event attributes.

Previous work in natural language inference has focused on linguistic entailment (Bowman et al., 2015; Bos and Markert, 2005) while ours focuses on commonsense-based inference. There also has been inference or entailment work that is more generation focused: generating, e.g., entailed statements (Zhang et al., 2017; Blouw and Eliasmith, 2018), explanations of causality (Kang et al., 2017), or paraphrases (Dong et al., 2017). Our work also aims at generating inferences from sentences; however, our models infer implicit information about mental states and causality, which has not been studied by most previous systems.

Also related are commonsense knowledge bases (Espinosa and Lieberman, 2005; Speer and Havasi, 2012). Our work complements these existing resources by providing commonsense relations that are relatively less populated in previous work. For instance, ConceptNet contains only 25% of our events, and only 12% have relations that resemble intent and reaction. We present a more detailed comparison with ConceptNet in Appendix C.

## 7 Conclusion

We introduced a new corpus, task, and model for performing commonsense inference on textually-described everyday events, focusing on stereotypical intents and reactions of people involved in the events. Our corpus supports learning representations over a diverse range of events and reasoning about the likely intents and reactions of previously unseen events. We also demonstrate that such inference can help reveal implicit gender bias in movie scripts.

## Acknowledgments

We thank the anonymous reviewers for their insightful comments. We also thank xlab members at the University of Washington, Martha Palmer, Tim O'Gorman, Susan Windisch Brown, Ghazaleh Kazeminejad as well as other members at the University of Colorado at Boulder for many helpful comments for our development of the annotation pipeline. This work was supported in part by National Science Foundation Graduate Research Fellowship Program under grant DGE-1256082, NSF grant IIS-1714566, and the DARPA CwC program through ARO (W911NF-15-1-0543).

## References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Apoorv Agarwal, Jiehan Zheng, Shruti Kamath, Sriramkumar Balasubramanian, and Shirin Ann Dey. 2015. Key female characters in film have more to

talk about besides men: Automating the bechdel test. In *NAACL*, pages 830–840, Denver, Colorado. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *COLING-ACL*.

Elizabeth Behm-Morawitz and Dana E Mastro. 2008. Mean girls? The influence of gender portrayals in teen movies on emerging adults' gender-based attitudes and beliefs. *Journalism & Mass Communication Quarterly*, 85(1):131–146.

Peter Blouw and Chris Eliasmith. 2018. Using neural networks to generate inferential roles for natural language. *Frontiers in Psychology*, 8:2335.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with robust logical inference. In *MLCW*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *CoNLL*.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *SSST@EMNLP*.

Yoonjung Choi and Janyce Wiebe. 2014. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In *EMNLP*.

Rebecca L Collins. 2011. Content analysis of gender roles in media: Where are we now and where should we go? *Sex Roles*, 64(3-4):290–298.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190. Springer.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

Haibo Ding, Tianyu Jiang, and Ellen Riloff. 2017. Why is an event affective? Classifying affective events based on human needs. In *AAAI Workshop on Affective Content Analysis*.

Haibo Ding and Ellen Riloff. 2016. Acquiring knowledge of affective events from blogs using label propagation. In *AAAI*.

Haibo Ding and Ellen Riloff. 2018. Weakly supervised induction of affective events by optimizing semantic consistency. In *AAAI*.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *EMNLP*.

Dawn Elizabeth England, Lara Descartes, and Melissa A Collier-Meek. 2011. Gender role portrayal and the Disney princesses. *Sex roles*, 64(7-8):555–567.

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

Vindu Goel and Mike Isaac. 2016. Facebook Moves to Ban Private Gun Sales on its Site and Instagram. https://www.nytimes.com/2016/01/30/technology/facebook-gun-sales-ban.html. Accessed: 2018-02-19.

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

Andrew S Gordon and Reid Swanson. 2008. StoryUpgrade: finding stories in internet weblogs. In *ICWSM*.

Philip John Gorinski and Mirella Lapata. 2015. Movie script summarization as graph-based scene extraction. In *NAACL*, pages 1066–1076.

Stephan Charles Greene. 2007. Spin: Lexical semantics, transitivity, and the identification of implicit sentiment.

Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. 2017. A question answering approach for emotion cause extraction. In *EMNLP*.

Joshua K. Hartshorne, Claire Bonial, and Martha Palmer. 2013. The verbcorner project: Toward an empirically-based semantic decomposition of verbs. In *EMNLP*.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Dongyeop Kang, Varun Gangal, Ang Lu, Zheng Chen, and Eduard H. Hovy. 2017. Detecting and explaining causes from text for a time series event. In *EMNLP*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *ACL*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Xiang Li, Aynaz Taheri, Lifu Tu, and Kevin Gimpel. 2016. Commonsense knowledge base completion. In *ACL*.

Nishtha Madaan, Sameep Mehta, Taneea S. Agrawaal, Vrinda Malhotra, Aditi Aggarwal, and Mayank Saxena. 2017. Analyzing gender stereotyping in bollywood movies. *CoRR*, abs/1710.04117.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38:301–333.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *ACL*.

Deborah A Prentice and Erica Carranza. 2002. What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of women quarterly*, 26(4):269–281.

Anil Ramakrishna, Victor R Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *ACL*, pages 1669–1678, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *ACL*.

Lena Reed, JiaQi Wu, Shereen Oraby, Pranav Anand, and Marilyn A. Walker. 2017. Learning lexico-functional patterns for first-person affect. In *ACL*.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *TACL*, 3:475–488.

Irene Russo, Tommaso Caselli, and Carlo Strapparava. 2015. Semeval-2015 task 9: Clipeval implicit polarity of events. In *SemEval@NAACL-HLT*.

Maarten Sap, Marcella Cindy Prasetio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *EMNLP*, pages 2329–2334.

Karin Kipper Schuler, Anna Korhonen, and Susan Windisch Brown. 2009. Verbnet overview, extensions, mappings and applications. In *NAACL*.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*.

Yla R Tausczik and James W Pennebaker. 2016. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.*

Emmett Tomai and Ken Forbus. 2010. Using narrative functions as a heuristic for relevance in story understanding. In *Proceedings of the Intelligent Narrative Technologies III Workshop*, page 9. ACM.

Hoa Trong Vu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Acquiring a dictionary of emotion-provoking events. In *EACL*.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In *EMNLP*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *TACL*, 3:345–358.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *TACL*, 5:379–395.

# A Appendix

## A.1 Event Extraction

We balance the number of content words to ensure that the events are generalizable but still concrete enough to be labelled. We only keep events with at least two and less than five content words, defined as words that are not stop words, person tags, or blanks. We count phrasal verbs (such as "get up") as content word. We limit the sets of events to those events that occur most frequently in our corpora, using corpus-specific thresholds. [5]

## A.2 Annotation Setup

Each event was presented to three different raters recruited via Amazon Mechanical Turk. Raters were given the option to say that the event did not make sense (invalid), at which point they were not asked any other questions. If the rater marked the event as valid, they were required to answer the question about how PersonX typically feels after the event. Each rater was paid $0.10 per event. Additionally we annotated a small number of events where "It" was in the subject (e.g., `It rains all day`). For these events, we only asked raters to say how other people typically feel after the event (if they marked the event as valid).

# B Event2Mind Training Details

In our experiments, we use Adam to train for ten epochs, as implemented in Tensorflow (Abadi et al., 2015).

For baseline models, the dimension of the event encoded embedding is $H = 300$. For our BiRNN model, we also experimented with an embedding dimension of $H = 100$.

We define the vocabulary as the tokens appearing in the training data events and annotations at least twice, plus the bigrams and trigrams that appear more than five times. In cases where an annotation for the intent/reaction was left blank (because there was no intent or the event did not affect other people), we treated the annotation as equivalent to the word "none". Because many of the annotations for intent started with "to" or "to be", we stripped these two words from the beginning of all intent annotations.

[5] For ROC Story and Spinn3r events, we choose events with frequency at least five and 100, respectively. For Syntactic Ngrams, we took the top 10000 events.

| Overlap criterion | % of Event2Mind events |
|---|---|
| Any node | 25% |
| All annotations, with select relations | 12 % |
| XIntent, with select relations | 3% |
| XReact/OReact, with select relations | <1% |

Table 5: Event2Mind events overlap with ConceptNet events. While a non-trivial amount are represented in some capacity, few events have intent or reactions.

# C Comparison with ConceptNet

We match our events with the event nodes in ConceptNet and find 6 ConceptNet relations that compare to our intent and reaction dimensions. Specifically, we compare *MotivatedByGoal, CausesDesire, HasFirstSubevent,* and *HasSubevent* with the 'XIntent' annotations, and 'XReact' and 'OReact' annotations with the *Causes* and *HasLastSubevent* relations. For each ConceptNet event, we then compute unigram overlap between our annotations and their ConceptNet proxy using the 6 relations.

We summarize overlap in Table 5, where we show that 75% of Event2Mind events are not covered in ConceptNet. We also show that while 12% of our events have an edge with one of the 6 relations, the actual overlap between our annotations and the ConceptNet data is very low (<5%). This overlap statistics indicates that our dataset provides new commonsense knowledge that is not covered by previous resources such as ConceptNet.

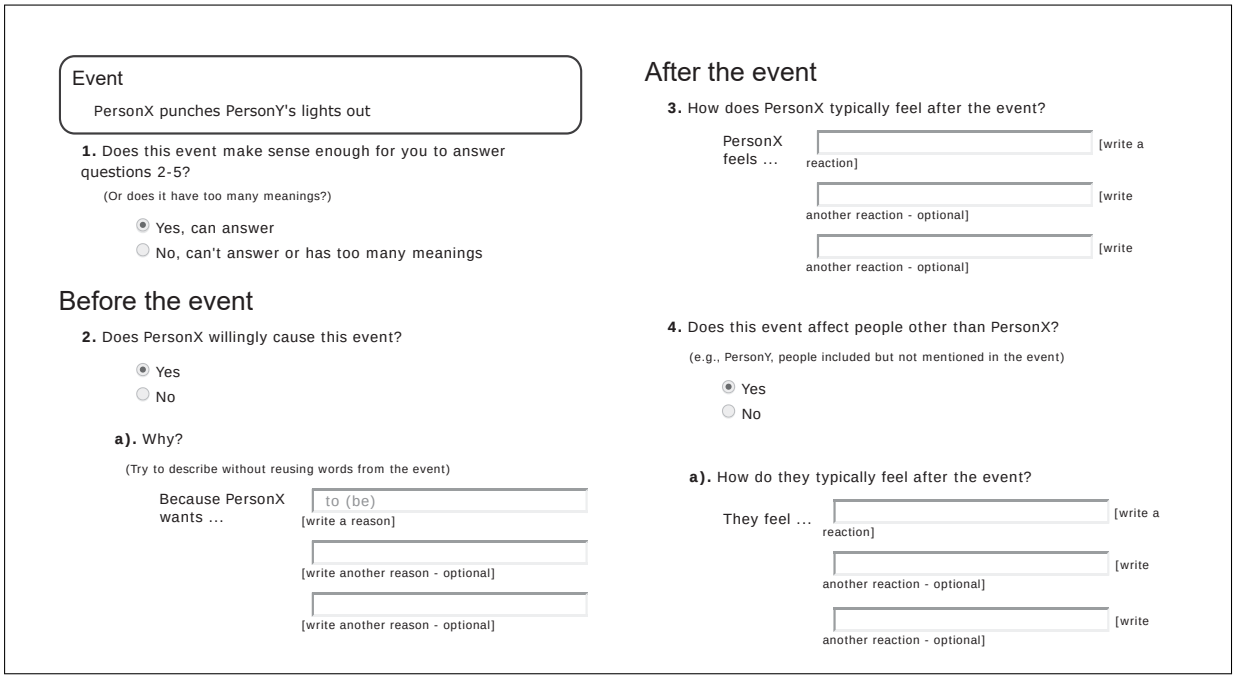


Figure 8: Main event phrase annotation setup. Each event was annotated by three Amazon Mechanical Turk raters.