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

Abstract

Understanding a narrative requires reading between the lines and reasoning about the unspoken but obvious implications about events and people's mental states — a capability that is trivial for humans but remarkably hard for machines. To facilitate research addressing this challenge, we introduce a new annotation framework to explain naive psychology of story characters as fully-specified chains of mental states with respect to *motivations* and *emotional reactions*. Our work presents a new large-scale dataset with rich low-level annotations and establishes baseline performance on several new tasks, suggesting avenues for future research.

1 Introduction

Understanding a story requires reasoning about the causal links between the events in the story and the mental states of the characters, even when those relationships are not explicitly stated. As shown by the commonsense story cloze shared task (Mostafazadeh et al., 2017), this reasoning is remarkably hard for both statistical and neural machine readers – despite being trivial for humans. This stark performance gap between humans and machines is not surprising as most powerful language models have been designed to effectively learn local fluency patterns. Consequently, they generally lack the ability to abstract away from surface patterns in text to model more complex implied dynamics, such as intuiting characters' mental states or predicting their plausible next actions.

In this paper, we construct a new annotation formalism to densely label commonsense short stories (Mostafazadeh et al., 2016) in terms of the mental states of the characters. The resulting dataset offers three unique properties. First, as highlighted in Figure 1, the dataset provides a fully-specified chain of *motivations* and *emotional reactions* for each story character as pre- and post-conditions of events. Second, the annotations include state changes for entities even when they are not mentioned directly in a sentence (e.g., in the fourth sentence in Figure 1, players would feel *afraid* as a result of the instructor throwing a chair), thereby capturing implied effects unstated in the story. Finally, the annotations encompass both formal labels from multiple theories of psychology (Maslow, 1943; Reiss, 2004; Plutchik, 1980) as well as open text descriptions of motivations and emotions, providing a comprehensive mapping between open text explanations and label categories (e.g., "to spend time with her son"

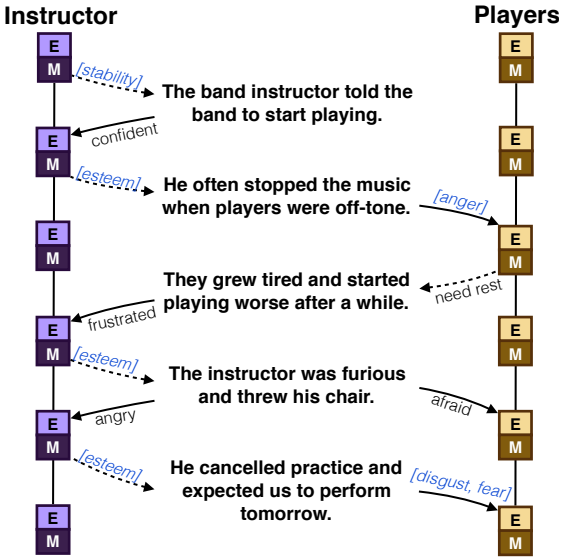

Figure 1: A story example with partial annotations for motivations (dashed) and emotional reactions (solid). Open text explanations are in black (e.g., "frustrated") and formal theory labels are in blue with brackets (e.g., "[esteem]").

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 2289–2299
Melbourne, Australia, July 15 - 20, 2018.

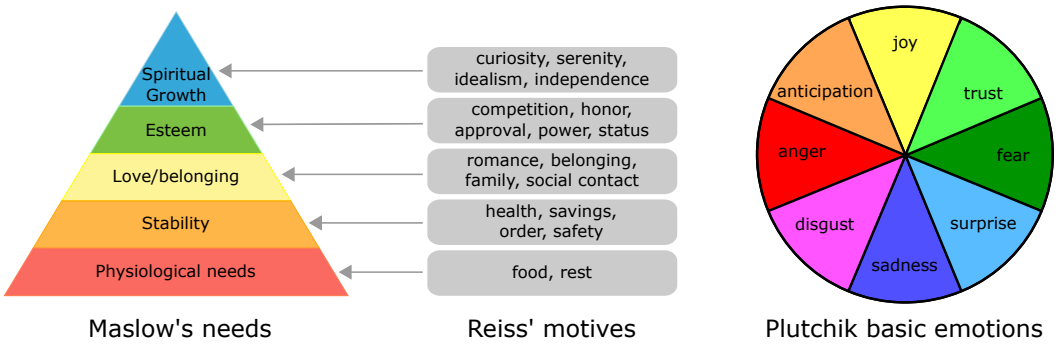

Figure 2: Theories of Motivation (Maslow and Reiss) and Emotional Reaction (Plutchik).

→ Maslow's category *love*). Our corpus[1] spans across 15k stories, amounting to 300k low-level annotations for around 150k character-line pairs.

Using our new corpus, we present baseline performance on two new tasks focusing on mental state tracking of story characters: *categorizing* motivations and emotional reactions using theory labels, as well as *describing* motivations and emotional reactions using open text. Empirical results demonstrate that existing neural network models including those with explicit or latent entity representations achieve promising results.

2 Mental State Representations

Understanding people's actions, motivations, and emotions has been a recurring research focus across several disciplines including philosophy and psychology (Schachter and Singer, 1962; Burke, 1969; Lazarus, 1991; Goldman, 2015). We draw from these prior works to derive a set of categorical labels for annotating the step-by-step causal dynamics between the mental states of story characters and the events they experience.

2.1 Motivation Theories

We use two popular theories of motivation: the "hierarchy of needs" of Maslow (1943) and the "basic motives" of Reiss (2004) to compile 5 coarse-grained and 19 fine-grained motivation categories, shown in Figure 2. Maslow's "hierarchy of needs" are comprised of five categories, ranging from *physiological needs* to *spiritual growth*, which we use as coarse-level categories. Reiss (2004) proposes 19 more fine-grained categories that provide a more informative range of motivations. For example, even though they both relate to the *physiological needs* Maslow category, the *food* and *rest* motives from Reiss (2004) are very different. While the Reiss theory allows for finer-grained annotations of motivation, the larger set of abstract concepts can be overwhelming for annotators. Motivated by Straker (2013), we design a hybrid approach, where Reiss labels are annotated as sub-categories of Maslow categories.

2.2 Emotion Theory

Among several theories of emotion, we work with the "wheel of emotions" of Plutchik (1980), as it has been a common choice in prior literature on emotion categorization (Mohammad and Turney, 2013; Zhou et al., 2016). We use the eight basic emotional dimensions as illustrated in Figure 2.

2.3 Mental State Explanations

In addition to the motivation and emotion categories derived from psychology theories, we also obtain open text descriptions of character mental states. These open text descriptions allow learning computational models that can *explain* the mental states of characters in natural language, which is likely to be more accessible and informative to end users than having theory categories alone. Collecting both theory categories and open text also allows us to learn the automatic mappings between the two, which generalizes the previous work of Mohammad and Turney (2013) on emotion category mappings.

3 Annotation Framework

In this study, we choose to annotate the simple commonsense stories introduced by Mostafazadeh et al. (2016). Despite their simplicity, these stories pose a significant challenge to natural language understanding models (Mostafazadeh et al., 2017).

[1]We make our dataset publicly available at https://uwnlp.github.io/storycommonsense/

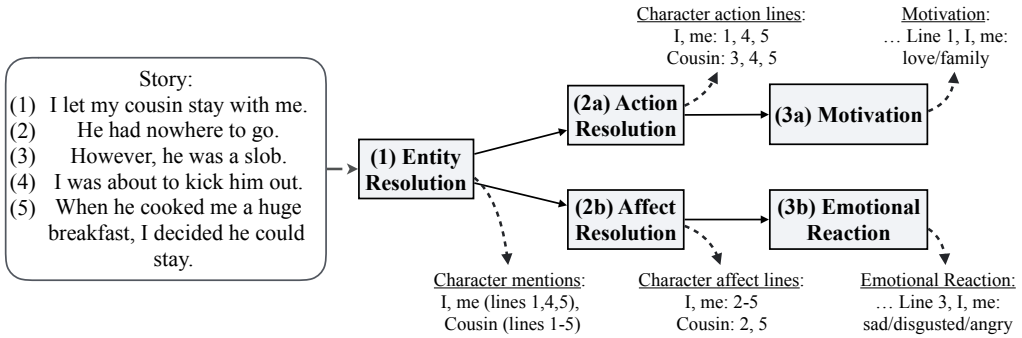

Figure 3: The annotation pipeline for the fine-grained annotations with an example story.

In addition, they depict multiple interactions between story characters, presenting rich opportunities to reason about character motivations and reactions. Furthermore, there are more than 98k such stories currently available covering a wide range of everyday scenarios.

Unique Challenges While there have been a variety of annotated resources developed on the related topics of sentiment analysis (Mohammad and Turney, 2013; Deng and Wiebe, 2015), entity tracking (Hoffart et al., 2011; Weston et al., 2015), and story understanding (Goyal et al., 2010; Ouyang and McKeown, 2015; Lukin et al., 2016), our study is the first to annotate the full chains of mental state effects for story characters. This poses several unique challenges as annotations require (1) interpreting discourse (2) understanding implicit causal effects, and (3) understanding formal psychology theory categories. In prior literature, annotations of this complexity have typically been performed by experts (Deng and Wiebe, 2015; Ouyang and McKeown, 2015). While reliable, these annotations are prohibitively expensive to scale up. Therefore, we introduce a new annotation framework that pipelines a set of smaller isolated tasks as illustrated in Figure 3. All annotations were collected using crowdsourced workers from Amazon Mechanical Turk.

3.1 Annotation Pipeline

We describe the components and workflow of the full annotation pipeline shown in Figure 3 below. The example story in the figure is used to illustrate the output of various steps in the pipeline (full annotations for this example are in the appendix).

(1) Entity Resolution The first task in the pipeline aims to discover (1) the set of characters E_i in each story i and (2) the set of sentences S_{ij} in which a specific character $j \in E_i$ is explicitly mentioned. For example, in the story in Figure 3, the characters identified by annotators are "I/me" and "My cousin", whom appear in sentences $\{1, 4, 5\}$ and $\{1, 2, 3, 4, 5\}$, respectively.

We use S_{ij} to control the workflow of later parts of the pipeline by pruning future tasks for sentences that are not tied to characters. Because S_{ij} is used to prune follow-up tasks, we take a high recall strategy to include all sentences that at least one annotator selected.

(2a) Action Resolution The next task identifies whether a character j appearing in a sentence k is taking any action to which a motivation can be attributed. We perform action resolution only for sentences $k \in S_{ij}$. In the running example, we would want to know that the cousin in line 2 is not doing anything intentional, allowing us to omit this line in the next pipeline stage (3a) where a character's motives are annotated. Description of state (e.g., "Alex is feeling blue") or passive event participation (e.g., "Alex trips") are not considered volitional acts for which the character may have an underlying motive. For each line and story character pair, we obtain 4 annotations. Because pairs can still be filtered out in the next stage of annotation, we select a generous threshold where only 2 annotators must vote that an intentional action took place for the sentence to be used as an input to the motivation annotation task (3a).

(2b) Affect Resolution This task aims to identify all of the lines where a story character j has an emotional reaction. Importantly, it is often possible to infer the emotional reaction of a character j even when the character does not explicitly appear in a sentence k. For instance, in Figure 3, we want to annotate the narrator's reaction to line 2 even though they are not mentioned because their emotional response is inferrable. We obtain 4 an-