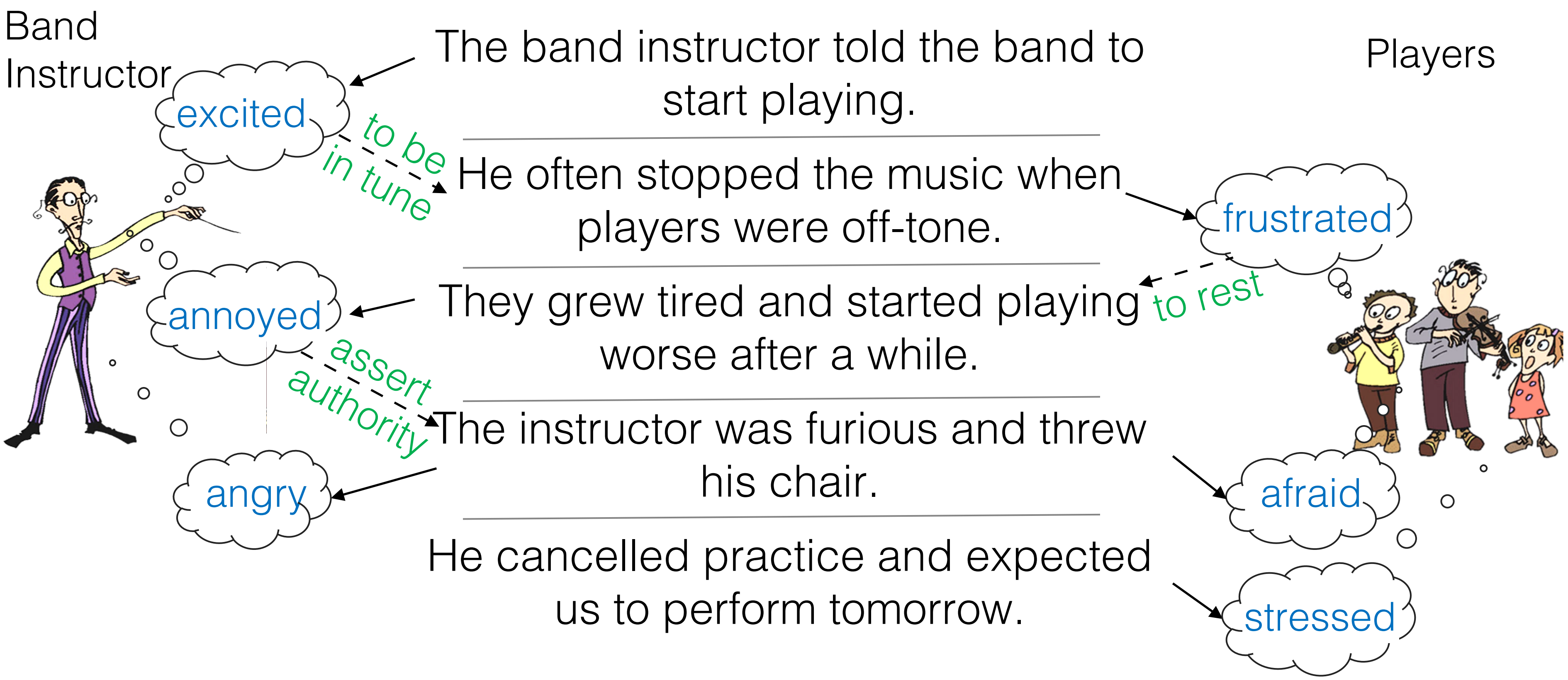
Modeling Naive Psychology of Characters in Simple Commonsense Stories

Hannah Rashkin, Antoine Bosselut, Maarten Sap,
Kevin Knight & Yejin Choi

Paul G. Allen School of Computer Science and Engineering, University of Washington
Allen Institute for Artificial Intelligence
Information Sciences Institute, University of Southern California

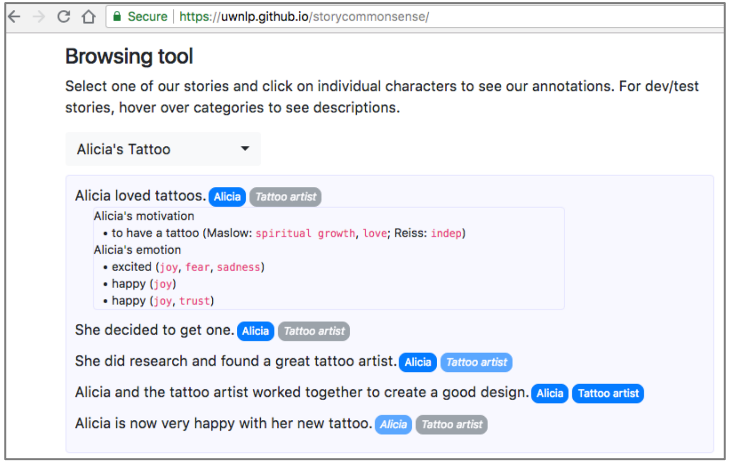
Inferring Character State



Reasoning about Naïve Psychology

New Story Commonsense Dataset:

- Open text + psychology theory
- Complete chains of mental states of characters
- Implied changes to characters
- Contextualized reasoning



https://uwnlp.github.io/storycommonsense/

How do we represent naïve psychology?

The band instructor told the band to start playing.
<u>He often stopped the music when players were off-tone.</u>

Psychology Theories

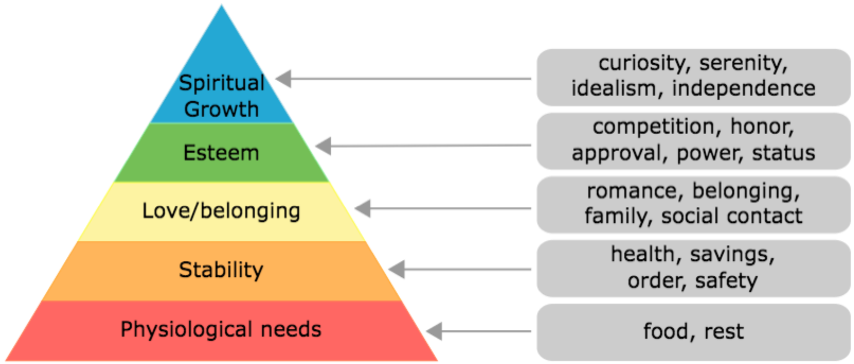
Natural Language

Instructor

Psychology Theories				Natural Language
Esteem	← wants	Instructor	wants →	To create a good harmony
Anger	← feels	Instructor	feels →	frustrated

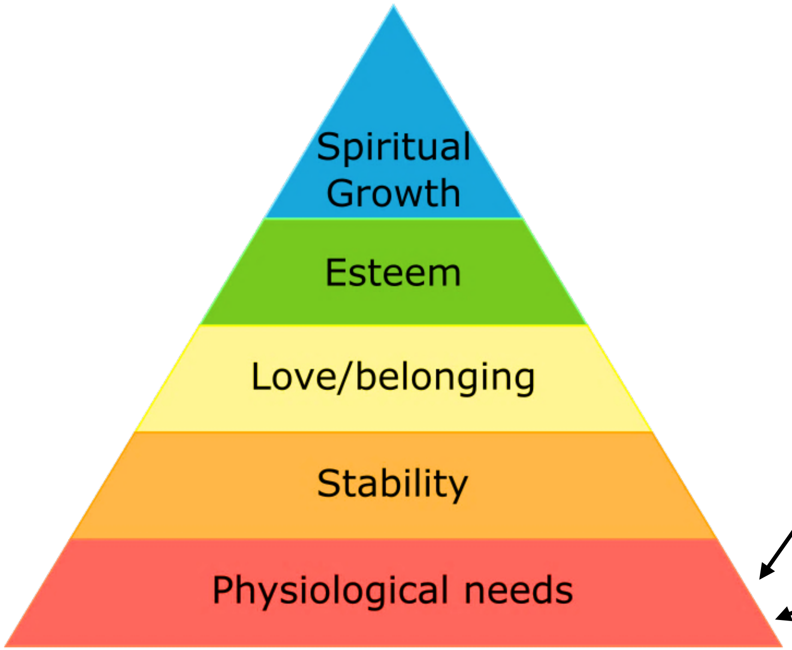
Naïve Psychology Annotations

- *Motivation:*
  - Causal source to actions
  - Motivational theories



- *Emotional Reaction:*
  - Causal effect of actions
  - Theories of emotion

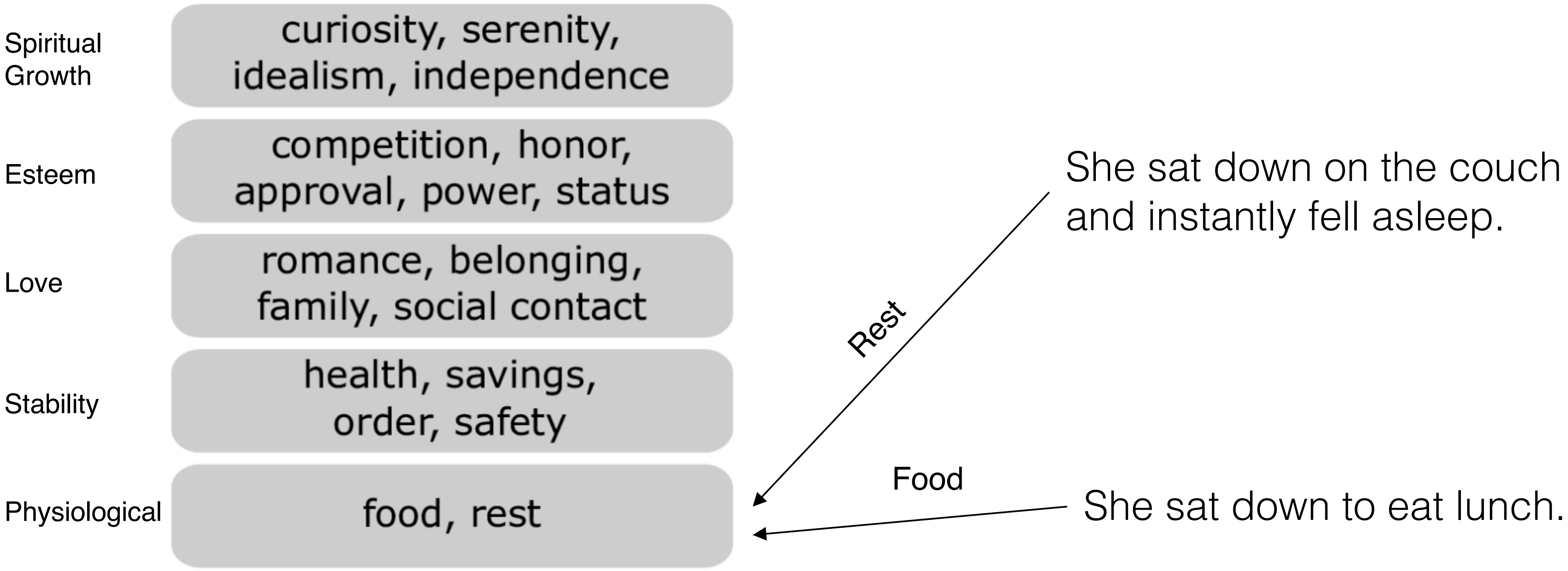
Motivation: Maslow Hierarchy of Needs (1943)



She sat down on the couch and instantly fell asleep.

She sat down to eat lunch.

Motivation: Reiss Categories (2004)



Emotional Reaction: Plutchik (1980)

Plutchik's Wheel

8 "main" emotions:

Their favorite uncle died.

↓

feel Sadness

Suddenly, they heard a loud noise.

↓

feel Fear, Surprise

Implicit Mental State Changes

The band instructor told the band to start playing.
He often stopped the music when players were off-tone.
They grew tired and started playing worse after a while.
<u>The instructor was furious and threw his chair.</u>

How are players affected?

→ implicitly involved

→ inference in these cases

Tracking Mental States

The band instructor told the band to start playing.
He often stopped the music when players were off-tone.
They grew tired and started playing worse after a while.
The instructor was furious and threw his chair.
<u>He cancelled practice and expected us to perform tomorrow.</u>

Why does the instructor cancel practice?

→ based on previous info

→ need to incorporate context

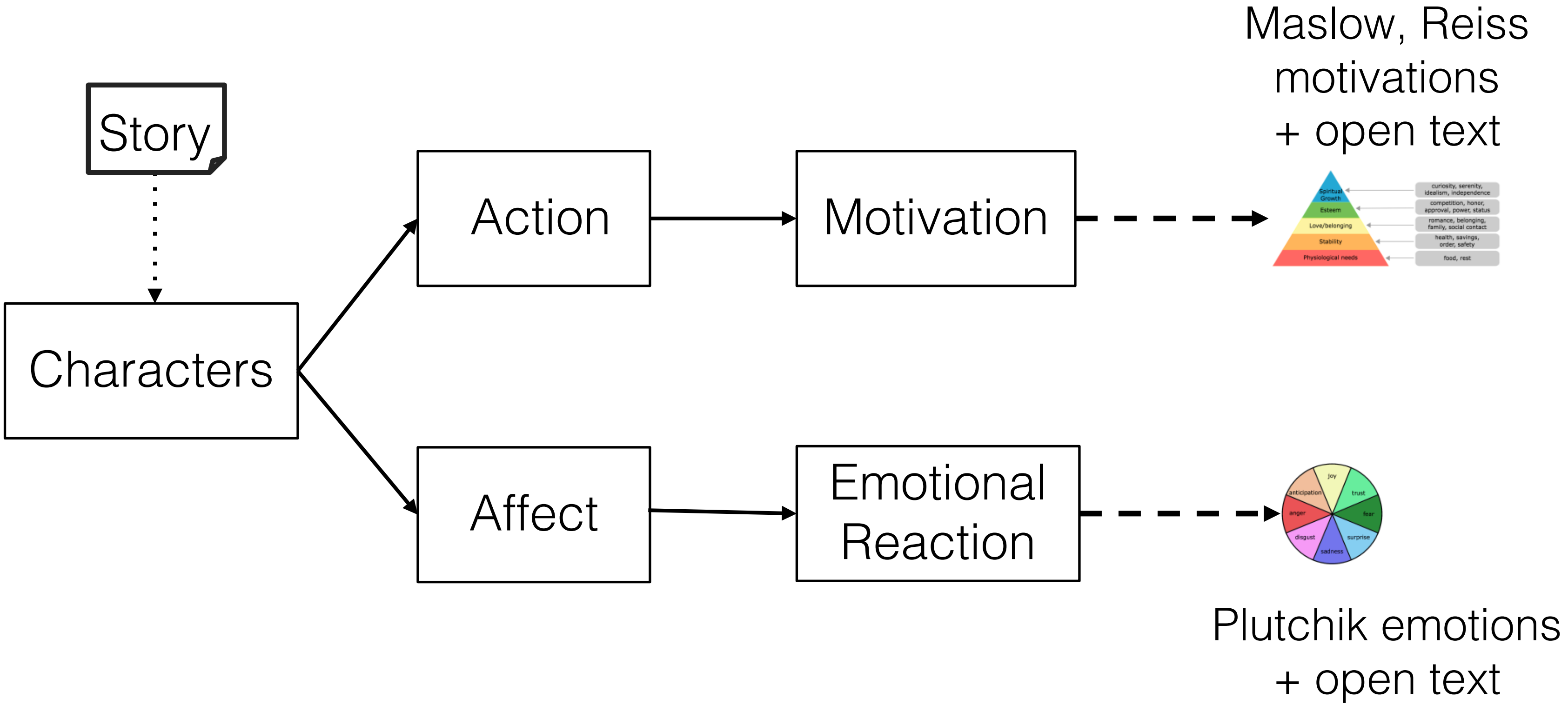
Related Work

- Reasoning about narratives (Mostafazadeh et al 2016)
- Detecting emotional content (Mohammad et al 2013) or stimuli (Gui et al 2017) of a statement

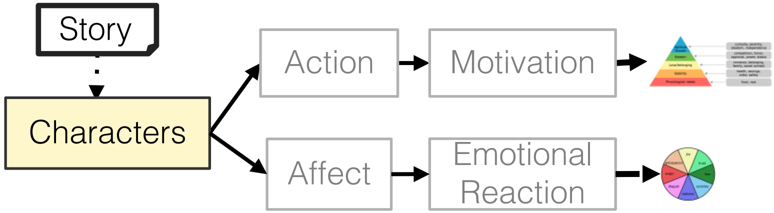
Our work:

- Both motivation and emotion for a character’s outlook
- Leverage psychology theories and natural language explanations

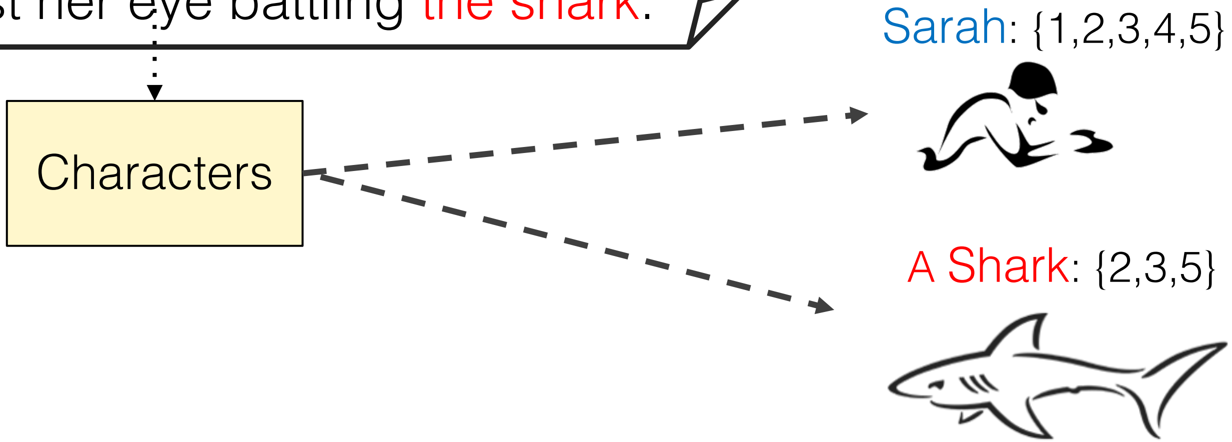
Full Annotation Chain



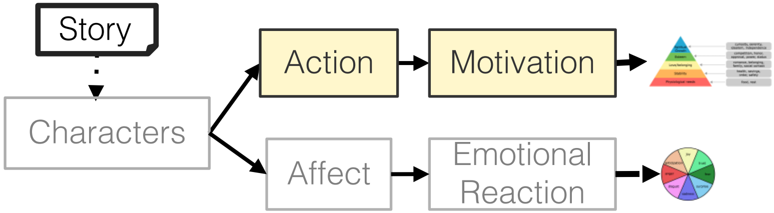
Full Annotation Chain



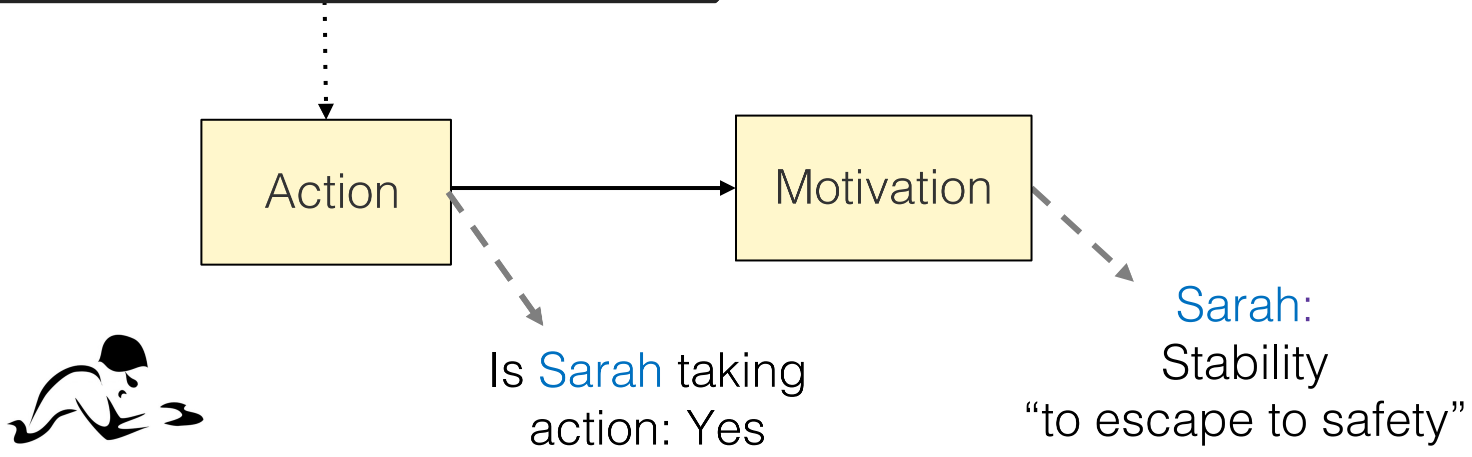
Sarah is swimming.
Sarah gets attacked by a shark.
Sarah fights off the shark.
Sarah escapes the attack.
Sarah lost her eye battling the shark.

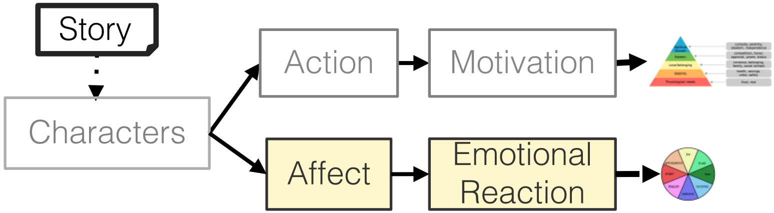


Full Annotation Chain



Sarah is swimming.
Sarah gets attacked by a shark.
Sarah fights off the shark.





Full Annotation Chain

Sarah is swimming.
Sarah gets attacked by a shark.
Sarah fights off the shark.

Affected

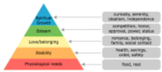
Emotional Reaction

Does the Shark have a reaction?
Yes

Shark:
Anger,
“aggressive”

Data Collection Summary

Over 300k low-level annotations for 15k stories from ROC training set

	Open-text	*Open-text + categories*		
	train	dev	test	
# character-line pairs	200k	25k	23k	
 … w/ motivation change	40k	9k	7k	*>50k motiv. changes*
 … w/ emotional reaction change	77k	15k	14k	*>100k emotion changes*

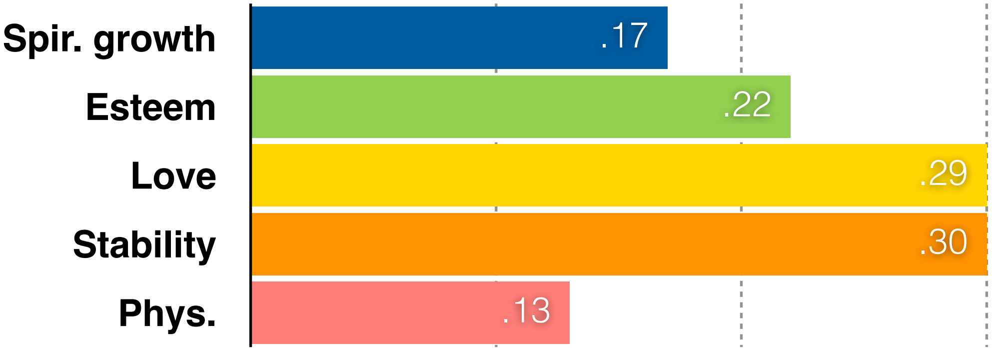
Annotated Data Distributions (Motivation)

- Fair amount of diversity in the open-text
- ~1/3 have positive motivation change:

Sampled Open-text Explanations

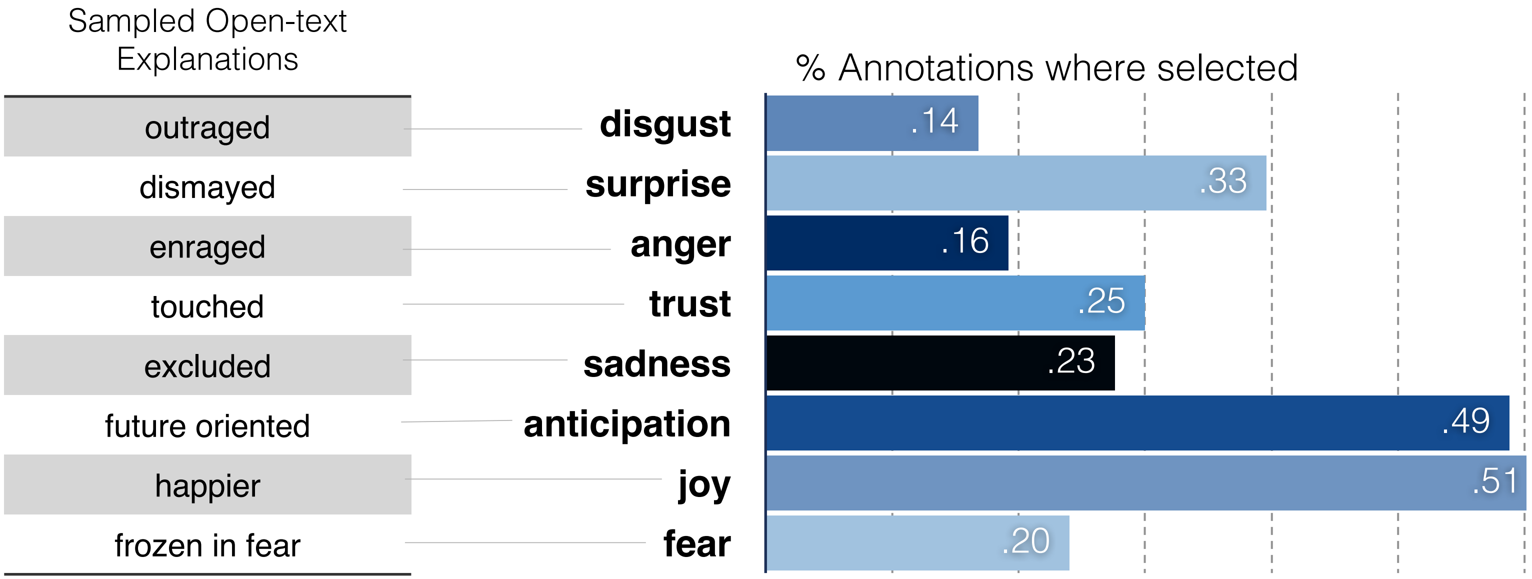
become experienced
meet goal; to look nice
to support his friends
be employed; stay dry
rest more; food

% Annotations where selected



Annotated Data Distributions (Emotion)

- Lots of happy stories
- ~2/3 have positive emotion change:



New Tasks

Given a story excerpt and a character can we explain the mental state:

- Explanation Generation: Generate open-text explanation of motivation/emotional reaction

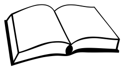
- State Classification: Predict Maslow/Reiss/Plutchik category

Task 1 - Explanation Generation

Explain mental state of character using natural language

Input

The band instructor told the band to start playing.



Story Text Excerpt + Character

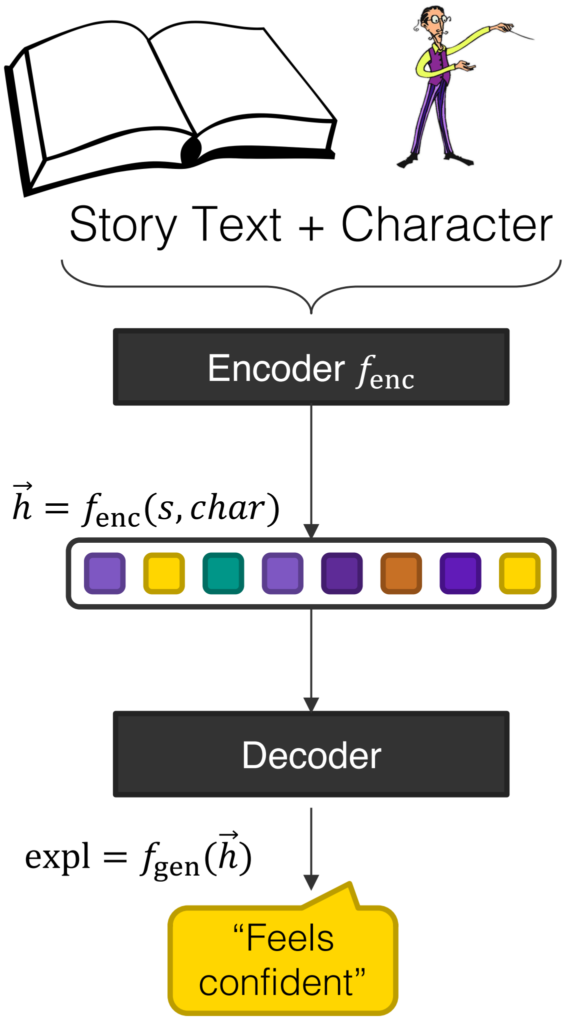
Output

“Feels confident”

Open Text Explanation

Modeling

- Using encoder-decoder framework
- Encoders - LSTM, CNN, REN, NPN
- Decoder for generation: single layer LSTM



Encoding Modules

Given entity e_j and line x^s (and entity-specific context sentences $x^c[e_j]$)

$$\boldsymbol{h} = \boldsymbol{f}_{\mathbf{enc}}\left(\boldsymbol{x}^s, \boldsymbol{x}^c[\boldsymbol{e}_j]\right)$$

Encoding functions:

- CNN, LSTM:
  encode last line and context -- concatenate

Entity Modeling

- Recurrent Entity Networks (Henaff et al 2017)
    - Store separate memory cells for each story character
    - Update after each sentence with sentence-based hidden states

- Neural Process Networks (Bosselut et al 2018)
    - Also has separate representations for each character
    - Updates after each sentence using learned action embeddings

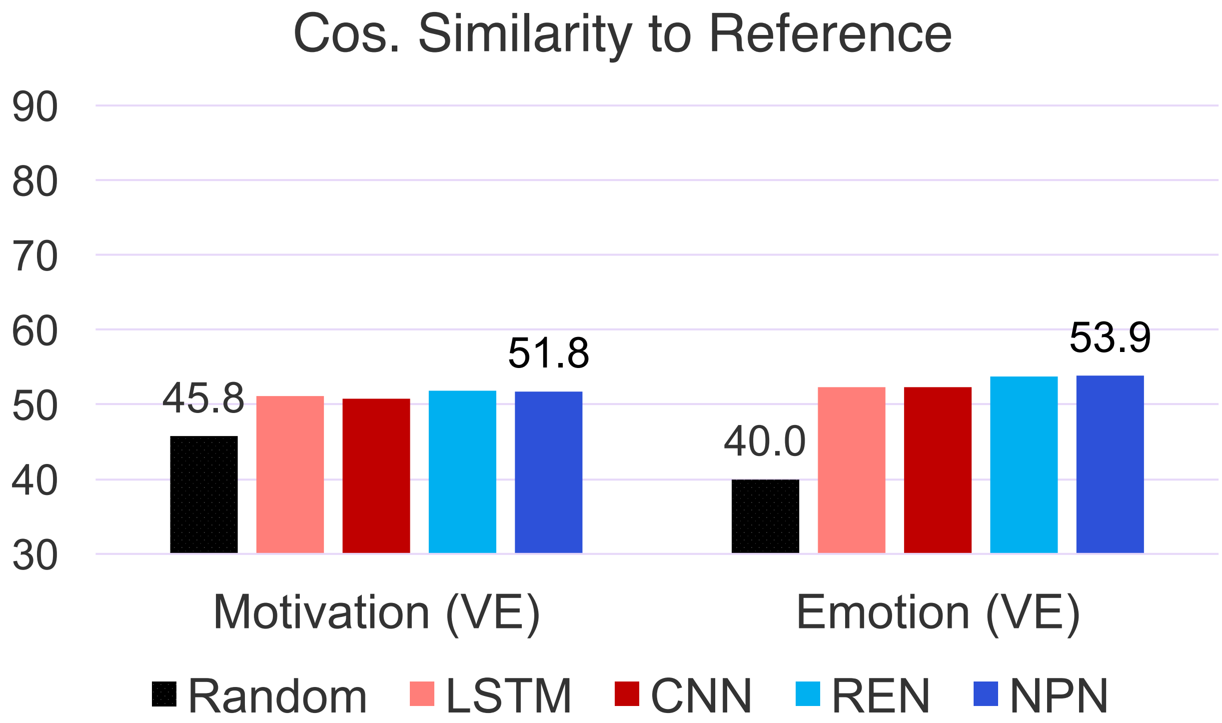
Explanation Generation Set-up

Evaluation: Cosine similarity of generated response to reference

Random baseline: Select random answer from dev set

- Responses are short/formulaic
- Words for describing intent/emotion are close in embedding space

Explanation Generation Results

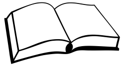


Task 2 – Mental State Classification

Predicting psychological categories for mental state

Input

The band instructor told the band to start playing.



Story Text Excerpt + Character

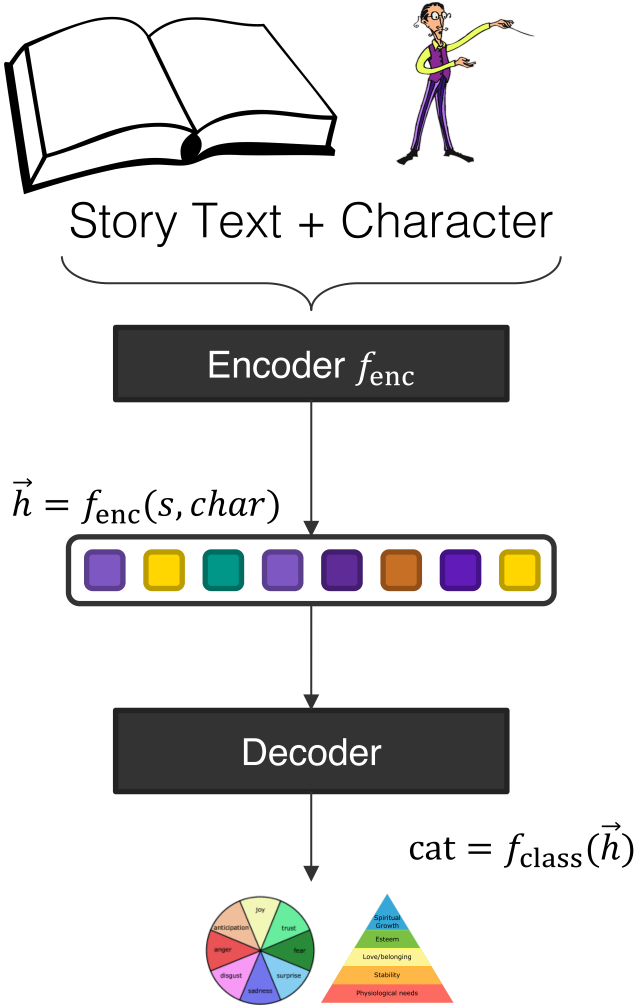
Output

anticipation

Theory categories

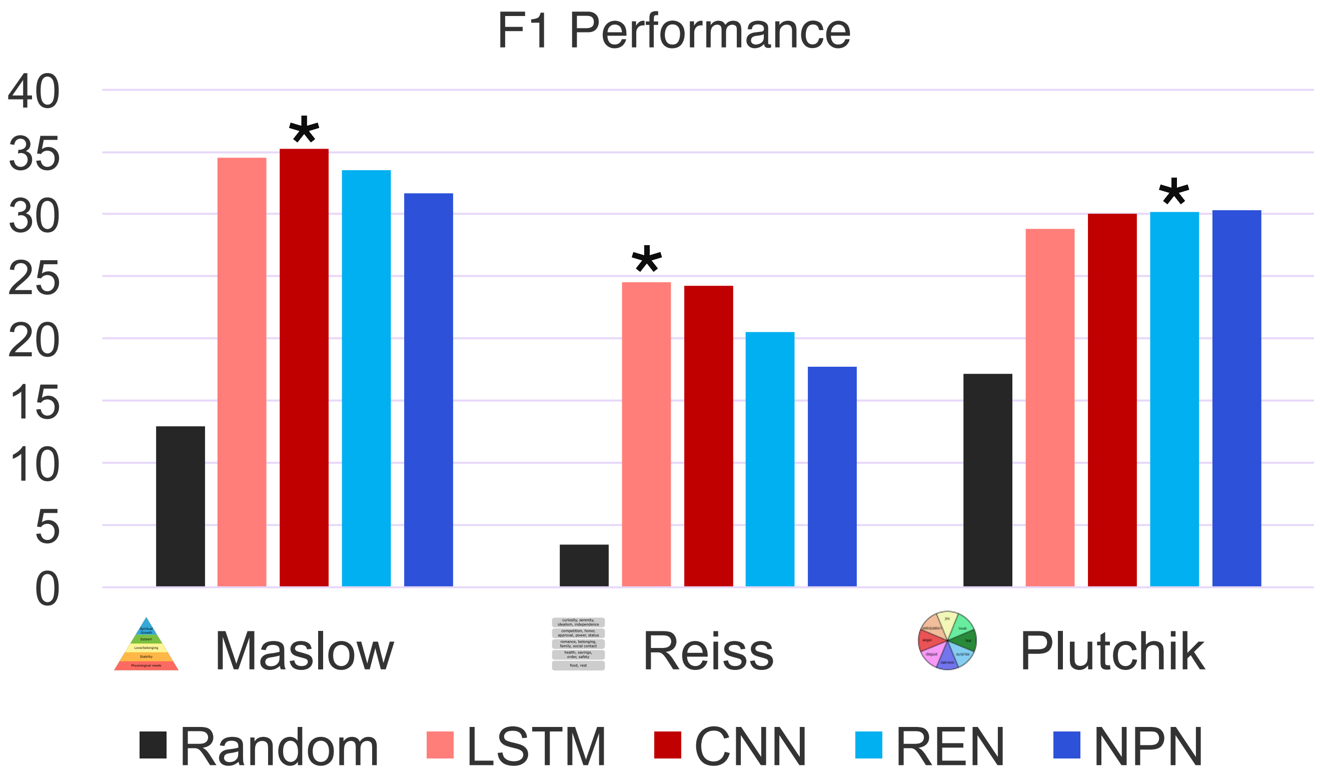
Modeling

- Using encoder-decoder framework
- Encoders - LSTM, CNN, REN, NPN
- Decoder for categorization: **logistic regression**



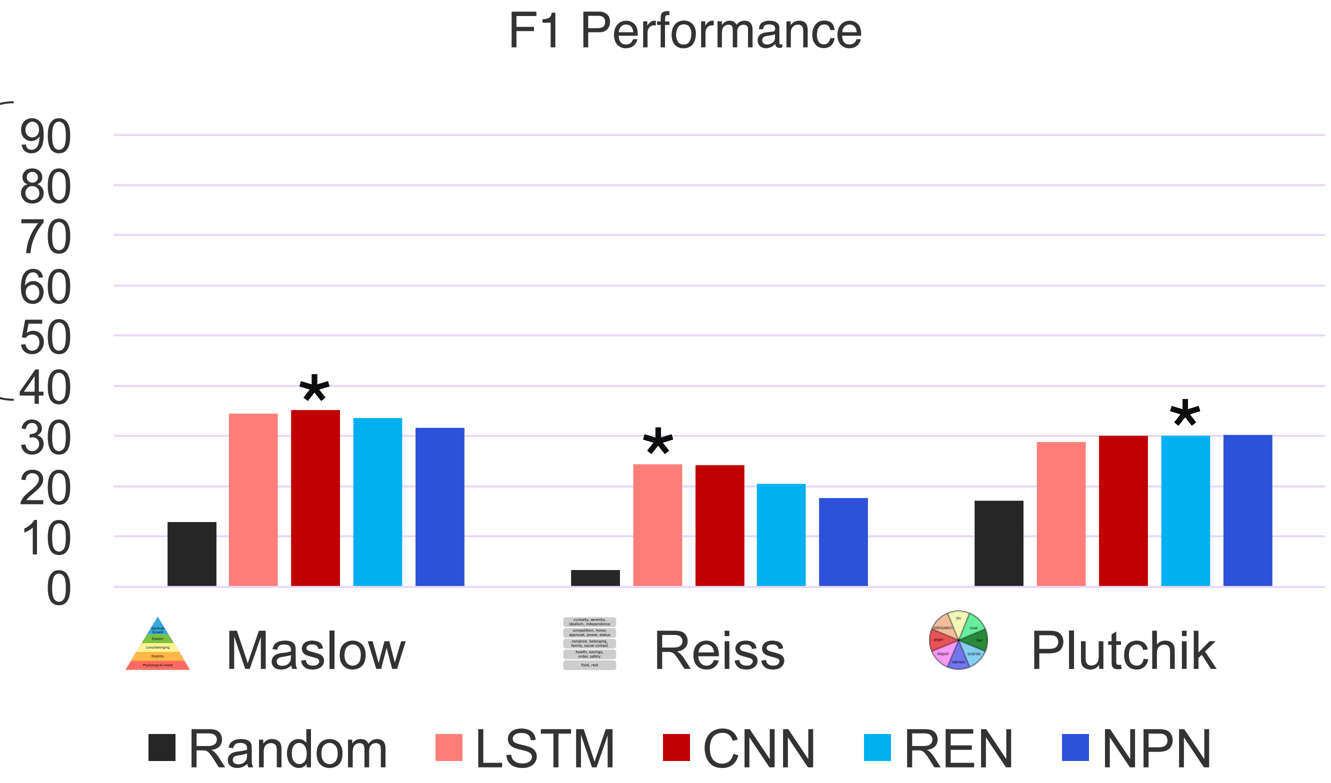
State Classification Results

- CNN and LSTM perform best on motivation categories
- Entity modeling has slight improvement in Plutchik



Further Improvement

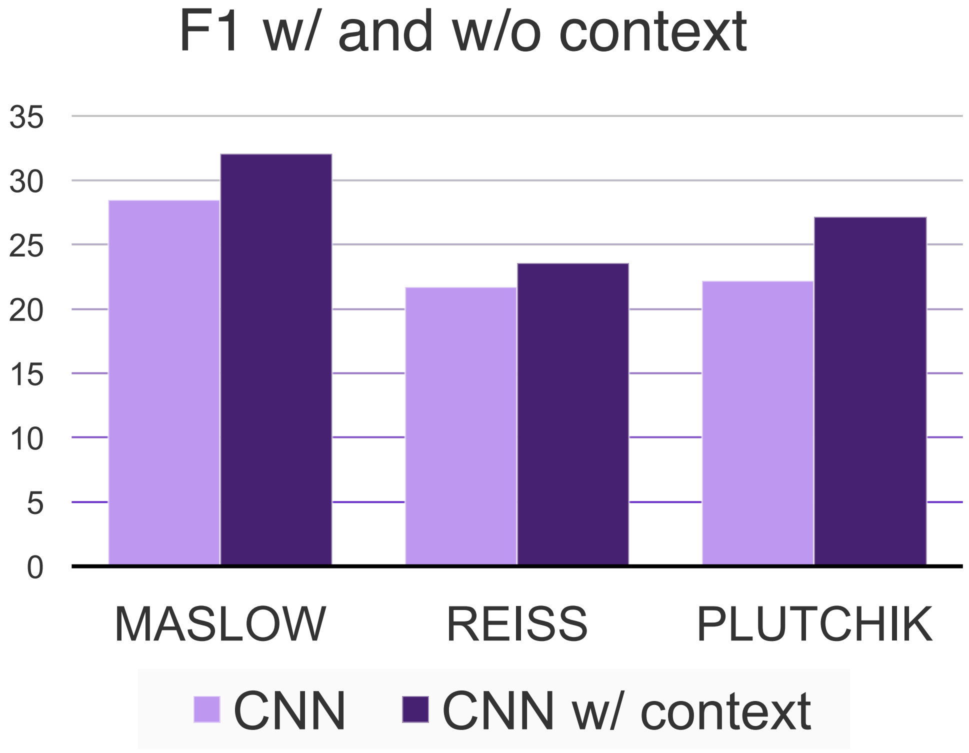
Best F1
at ~35%



Effect of Entity Specific Context

Including previous lines from context that include entity

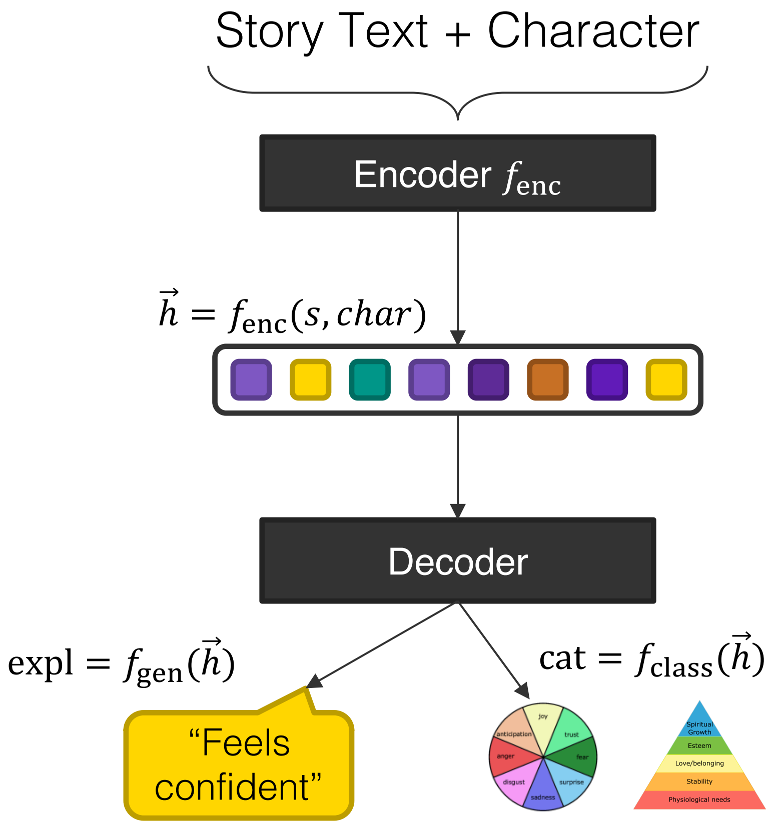
Entity specific context: improves all models F1 by about 3-5%



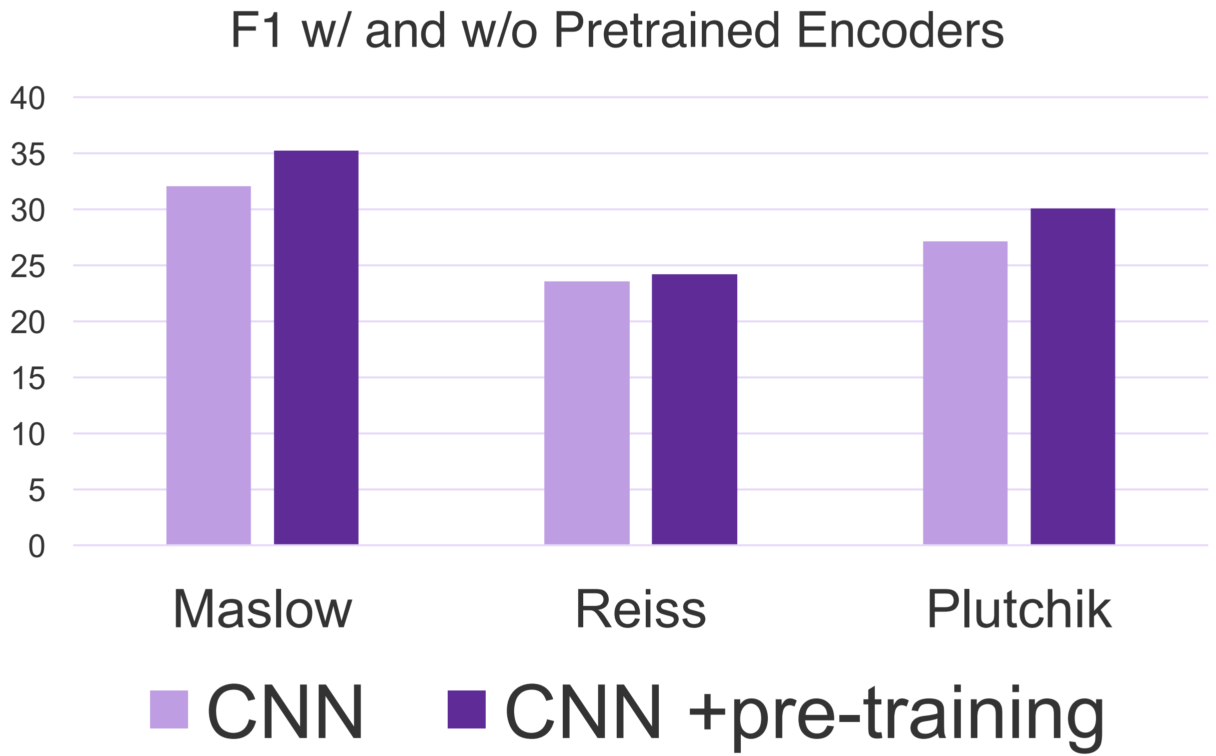
Pre-training Encoders

We have more open-text explanations than category annotations:

1. Pre-train encoders on open-text explanations
2. Fine-tune with the categorical labels



Effect of Pretrained Encoders

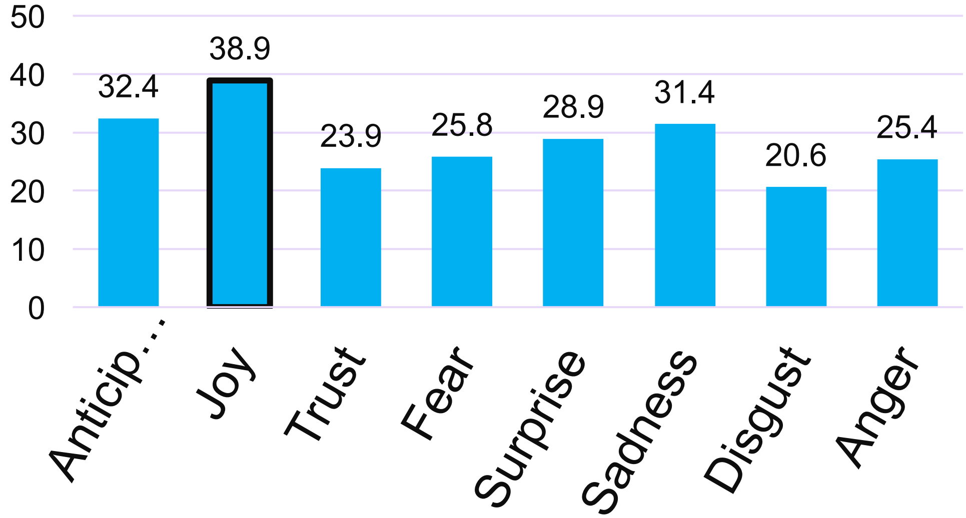


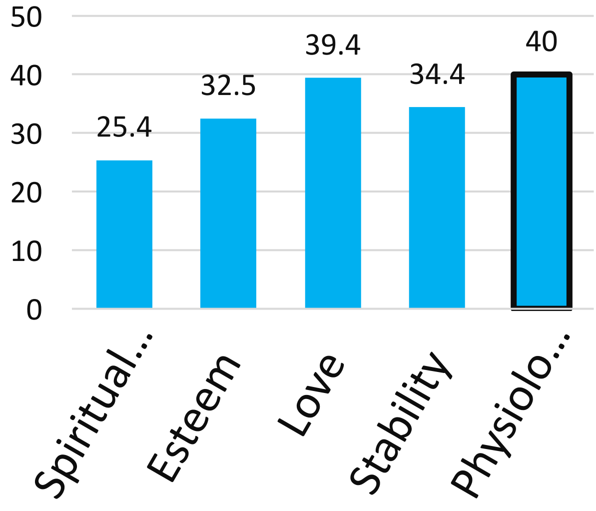
Improves:
1-2%

Performance Per Category

Highest performance:

- Frequent classes (eg. “joy” F1: 38.9%)
- Very concrete sets of actions (“physiological” F1: 40%)





Future Work

- **Outside Knowledge**: Help with infrequent classes and subtle implied changes

- **Social Commonsense**: Help with inferring mental state especially in more contextual cases

- **Potential Applications**: Improving language models, chat systems, natural language understanding

Conclusions

- New Dataset:
  - 15k roc stories annotated per character
    - >50k motivation changes
    - >100k emotions changes
  - `https://uwnlp.github.io/storycommonsense/`