Semantics with Feeling: Emotions for Abstract Embedding, Affect for Concrete Grounding

Daniele Moro
Computer Science
Boise State University
1910 W University Dr.
Boise, ID 83725
danielemoro@u.boisestate.edu

Gerardo Caracas
Computer Science
Boise State University
1910 W University Dr.
Boise, ID 83725
gerardocaracasur@u.boisestate.edu

David McNeill
Computer Science
Boise State University
1910 W University Dr.
Boise, ID 83725
davidmcneill@u.boisestate.edu

Casey Kennington
Computer Science
Boise State University
1910 W University Dr.
Boise, ID 83725
caseykennington@boisestate.edu

Abstract

An important yet underexplored aspect of meaning in both distributional and grounded models of semantics is emotion. In this paper, we explore how emotion can be predicted from descriptions of robot behaviors represented with embeddings. We then compare this approach with a grounded model that maps corresponding robot behaviors represented as internal states to the same emotion labels and discover comparable results. We then take the predictions from the second model and use them as a proxy for concrete affect (as opposed to abstract emotion) and use this derived affect to ground a semantic classifier in a retrieval task and see improvements on the retrieval task when affect is used as a grounded modality. This demonstrates that semantics can benefit from using a proxy of affect derived from human perceptions, given those perceptions are mapped to clear proxies, such the behaviors of an embodied robot.

1 Introduction

Semantic representation is a crucial part of language understanding for spoken dialogue systems, and the semantic meanings of many words have emotion as part of their connotative meaning (Lane and Nadel, 2002). Yet, including emotion in semantic models is complicated by the divide between distributional and grounded semantic representations, which rely upon separate assumptions for the source of meaning: distributional models only consider abstract meaning and grounded models only consider concrete meaning. Indeed, Bender and Koller (2020) and Bisk et al. (2020) observed that models of distributional semantics (i.e., embeddings) operate only on text and are missing key aspects of meaning. We therefore infer that distributional approaches make an *abstractness* assumption because the mode of acquisition for abstract language is other linguistic information (Della Rosa et al., 2010). Conversely, grounded semantic models make a *concreteness* assumption in that all semantic information can be acquired concretely by physical world denotations. Following Barrett (2017), emotions can similarly be dichotomized as abstract and concrete; abstract according their lexical categories (e.g., *happiness*, *fear*, *anger*) distributed with text, and concretely through *affect* which is a biological system and a fundamental part of embodiment (Vigliocco et al., 2014). In contrast to abstract emotion concepts, affect is a more basic underpinning for emotion, ranging from unpleasant to pleasant (valence) and from agitated to calm (arousal).

In this paper, we explore how emotion can be approached abstractly from embeddings derived from written descriptions describing a robot's behavior, and how affect can be approached concretely by deriving affect from the physical internal state data of a robot. We tie this to a semantic model by addressing the question of how a *grounded* semantic model could use a representation of affect and how this compares to existing work that links emotions to abstract semantic representations (like embeddings). In Experiment 1, we model abstract emotions from descriptions of robot behaviors in accordance with the abstractness assumption. In Experiment 2, we model concrete emotions from the robot's internal state data in accordance with the concreteness assumption. We hypothesize that both text descriptions and internal robot states can be used to classify emotion labels and conclude that these approaches are comparable. In Experiment 3, we take the model from Experiment 2 and use the distribution over emotion labels derived from internal robot states as a feature vector of affect for a grounded semantic model. The results show that the semantic model can ground into our representation of affect.

2 Related Work

Our work relates to other recent work in connecting abstract emotions with embeddings. Xu et al. (2018) introduced Emo2Vec, a model of representing emotion as an embedding using multi-task learning. Agrawal et al. (2018) enriched word representations with emotional information, taking into account the fact that emotion words are distributed in similar ways in text, but have vastly different underlying emotional affect (e.g., *sad* is distributionally similar to *happy*). Similarly, Saravia et al. (2018) introduced CARER, a semi-supervised approach for representing lexically contextualized affect, enriched with embeddings. Finally, Alhuzali et al. (2018) extracted emotional information in text using Arabic dialectic first-person seed words. Ongoing work in emotion representation continues to follow the distribution hypothesis (i.e., assuming the semantics of all words are abstract), comparable to our findings in Experiment 1, but we move beyond this work by exploring how words can ground into an extra-linguistic affective representation.

Also related to our work is research on grounding semantic meaning of words into perceptual modalities beyond vision, such as auditory (Kiela and Clark, 2015) and olfactory perception (Grabski et al., 2012), haptics (Alomari et al., 2017; Thomason et al., 2016), and simulated hand muscle activations (Moro and Kennington, 2018). More directly related to our work is Song and Yamada (2018), which reported a multimodal approach in predicting a human's label of robot affect from seven basic emotions, as well as overall valence and arousal. Our work adds to a growing literature around a notion of embodied semantics (Johnson, 2008; Goertzel et al., 2010)–our experiments show that abstract words can ground more readily into representations of affect and emotion.

Our work builds directly off of McNeill and Kennington (2019), which explored how humans interpret affective display of robot behaviors, and Novikova et al. (2015) which explored how dynamic behaviors can be represented for mapping to emotions. We extend their work by linking dynamic robot modalities to emotion labels, and also to text descriptions of those robot behaviors.

3 Models

To model language abstractly for Experiment 1, we use the BERT model introduced in Devlin et al. (2018), which has been leveraged in many language tasks, in many cases resulting in state-of-the-art performance.

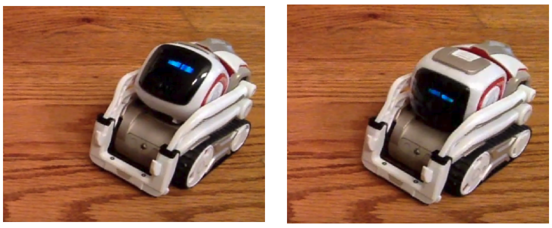

Figure 1: Two example frames of a recorded behavior.

To model grounded semantics for Experiment 3, we use the *words-as-classifier* (WAC) model for grounded lexical semantics (Kennington and Schlangen, 2015) due to its simplicity and interpretability. The WAC model uses a task-independent approach to predicting the semantic appropriateness of a word given a physical context. It pairs each word w in its vocabulary V with a classifier, and this classifier maps the real-valued features x of an experience exp (e.g., information recorded by a robot's sensors) to a semantic appropriateness score (i.e., if the information belongs to the class given by a word):

$$[\![w]\!]_{exp} = \lambda\mathbf{x}.p_w(\mathbf{x}) \tag{1}$$

For example, to learn the grounded meaning of the word *turn*, the low-level features (e.g., wheel motor speed) of a robot behavior described by the word *turn* are given as positive instances to a supervised learning classifier. Negative instances are randomly sampled from the complementary set of descriptions (i.e., behaviors not described by the word *turn*). This results in a trained $\lambda\mathbf{x}.p_{turn}(\mathbf{x})$, where x is the feature-set of a behavior that can be described by $turn$.

4 Data

For all of our experiments, we use data introduced in McNeill and Kennington (2019) that used the Anki Cozmo robot (see Figure 1).[1] Cozmo was marketed as a toy robot for children, is small in size, and the SDK allows developers to access low-level information about the robot state and to control several degrees of freedom including animated eyes, a head and lift that can move up and down, wheels that can be used to turn and move

[1] https://github.com/bsu-slim/cozmo-affect-data

the robot forward or backward, and a speech synthesizer. Cozmo has a built-in camera and simple built-in object and facial detection software. Following McNeill and Kennington (2019), we use the data they collected by asking Amazon Mechanical Turk workers to observe and write English descriptions and label of Cozmo's 941 prescripted behaviors for emotion. These behaviors include movements, sounds, and facial animations that are easily observable by a person.

Their data collection resulted in 1,870 descriptions and 16 **emotion labels** (*interest, alarm, confusion, understanding, frustration, relief, sorrow, joy, anger, gratitude, fear, hope, boredom, surprise, disgust, desire*), at least two for each of the behaviors (their work focused on the emotion labels; we use both the workers' emotion labels and their written descriptions). We normalized the descriptions by lower-casing all text, replacing punctuation with spaces, and removing any non-alphanumeric characters, resulting in a vocabulary size of 1230 words. The average phrase length was 7.9 words. The most common reported emotion label was `interest` (22% of phrases). The original dataset has three modalities: audio, facial, and internal states which represent Cozmo's animation modalities for each behavior. In this paper, we focus only on internal states (due to the feature transform we apply below). Each state update recorded a vector of 47 continuous feature values. The number of state updates varied for each behavior; i.e., a behavior could have as few as one state update vector, or as many as over a thousand state update vectors.

We augment their dataset with features explained in Novikova et al. (2015) (see Table 1 in that paper) which we term *Novikova* features, that we use in Experiments 2 & 3. These features are a functional transformation from the internal states, (which could range from a few state updates to as many as two thousand state updates for a particular behavior) to a set of 9 features:[2]

- Approach 1 - Transfer weight forward (head bent or movement forward)
- Approach 3 - Move its body forward (track wheel movement forward)
- Approach 5 - Extend or expand its body (lift movement up)
- Avoidance 6 - Transfer weight backward (head bent or movement backward)
- Avoidance 8 - Move its body backward (track wheel movement backward)
- Avoidance 9 - Attract limbs close to body (lift movement down)
- Energy 11 - High strength (high wheel speed)
- Energy 12 - Low strength (low wheel speed)
- Flow 18 - High change in tempo (change in motor speed)

Each feature yields a value that is a percentage of the time that feature is true. For example, Approach 1 is the percentage of robot state changes with forward movement and Avoidance 9 is the percentage of the state updates where the lift was in the lower half. Taken together, these transformations result in a vector of 9 values, each value between 1 and 0. Using these features has the added benefit of being generalizable to other robots; one only needs to map internal state representations of their chosen robotic platform to the Novikova features for modeling emotion.

Example Figure 1 shows two frames of a recorded behavior corresponding to Example (1) below, which includes (a) one worker's typed description of the behavior, (b) that worker's chosen affects to label the behavior, (c) a table with a sample of 4 internal features at one state update, and (d) Novikova features.

(1) a. The robot comes closer to the camera, showing a desire to know what is going on. Then it slams down its arms, like it wants to know.

b. sad, frustrated, curious

c.

feature	**value**
left wheel speed	0.0
right wheel speed	0.0
lift height	4.0
head position	3.0
...	

d. [0.11, 0.27, 0.97, 0.8, ..., 0.34]

Analysis of Novikova Features Figure 2 shows an application of TSNE to our data, which are represented as the Novikova features, annotated by color for each emotion. There are some clear clusters for some of the emotions, such as frustration, surprise, alarm, boredom, interest, and sorrow, indicating that many of the robot behaviors are cleanly separated, but there are many instances where a behavior does not belong to a particular cluster, indicating noise from human judgements of emotion of robot behaviors can be subjective, but potentially useful if a classifier can pick up on differences, which we explore in Experiment 2.

[2] We excluded the other 14 features because they did not change the results for a subset of Experiment 2; i.e., for our data, they were interpolations of the 9 remaining features.

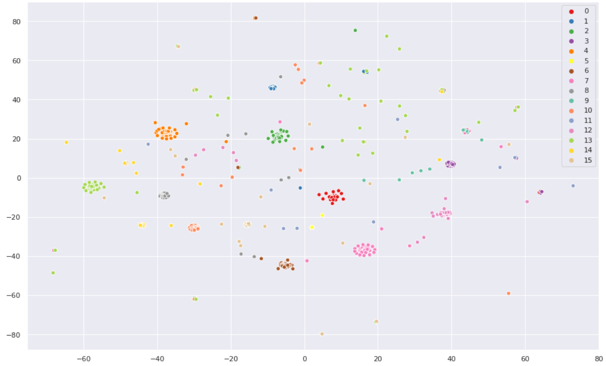

Figure 2: TSNE applied to all robot behaviors represented by the Novikova features colored for each of the 16 emotion labels. 0=interest, 1=alarm, 2=confusion, 3=understanding, 4=frustration, 5=relief, 6=sorrow, 7=joy, 8=anger, 9=gratitude, 10=fear, 11=hope, 12=boredom, 13=surprise, 14=disgust, 15=desire

Accuracy	Precision	Recall	F1 Score
90.4%	69.6%	31.4%	43.3%

Table 1: Results from Experiment 1 model when predicting 16 binary emotion labels on the testing set using a threshold of 0.5.

5 Experiments

5.1 Experiment 1: Classifying Emotion Labels from Descriptions

In this experiment, we explore how language maps to abstract emotions; that is, by using the BERT model we are making the abstractness assumption on the descriptions, and we are assuming that emotions are labels of abstract concepts. As has been argued by Barrett (2017), emotions are indeed abstract categorical labels that are derived from affect and are tied into the abstract linguistic system. By mapping embeddings to a distribution over 16 emotion labels, which themselves are a representation of the human judgment of emotion, we can ascertain if emotion is a suitable modality that abstract language can map to.

Task & Procedure For this experiment, we evaluate a model that takes as input a written description of the robot behaviors, and outputs a distribution over the 16 emotion labels. The data was split into training and testing groups using a random 90:10 split.

Model and Training We used the BERT model introduced by Devlin et al. (2018), as BERT has been shown to improve over traditional recurrent models. Instead of fine-tuning the pre-trained BERT model, we use a feature-based approach that has been shown to approximate fine-tuning, where "fixed features are extracted from the pre-trained model" (Devlin et al., 2018). The embeddings extracted from the pre-trained BERT transformer encoder (pre-trained on BooksCorpus (800M words) (Zhu et al., 2015) and English Wikipedia (2,500M words)) are padded to a length of 80, then fed into a two-layer feedforward neural network consisting of 128 nodes followed by 16 nodes (the small network is a result of our small amount of data). We trained using a binary crossentropy loss function and the adam optimizer, for a period of over 800 epochs (hyperparameters were chosen empirically using a subset of the training data).

Metrics We report accuracy, precision, recall, and F1 score of the trained model, assigning a probability above 0.5 to each label (i.e., more than one label could be correct) using a micro averaging strategy.

Results The results for this experiment are displayed in Table 1. These results show that our model can effectively map a description of a robot action to a distribution over emotion labels with relatively little data. Analysis of a confusion matrix shows that most of the errors are presented in the form of *false positives*. This could be caused by multiple users labeling the same robot action with conflicting labels.

5.1.1 Analyses

Here we report further analyses to determine what the model in this experiment is learning.

Synthetic Descriptions To explore whether the model predicts what we would expect, Table 2 shows synthetic descriptions that we applied to the model and observed the affect with the highest probability as assigned by the model.

Description	Predicted Emotion
the robot looks down	fear
the robot looks up	interest
the robot moves away	boredom
the robot moves closer	interest
the robot shakes his head	frustration
the robot nods his head	understanding
the robot squeals	gratitude
squinting eyes	confusion

Table 2: Examples of predicting emotion labels from descriptions of robot actions.

Word-Emotion Associations Focusing on the word-level, we analyze what our model is learning by determining which words correspond to specific emotion labels (i.e., we applied single words of our vocabulary to the model and observed the resulting distribution over the emotion labels). Table 3 shows several noteworthy examples of words with their corresponding top three ranking emotion labels in the model's resulting distribution after applying that word. As expected, *sorrow*'s most probable emotion label is sorrow (as was also the case for *understanding*, *confusion* and other emotion words), but other words that aren't directly related to emotion label also showed strong predictions. The word *eyes* corresponds to confusion and alarm (i.e., eyes furrow or widen, respectively) and when the robot lowers its lift, it is interpreted as sorrow (i.e., disappointment) or desire (i.e., lowering the lift keeps the lift out of the robot's camera view so it can better observe objects).

Word	Top Emotion Labels
sorrow	sorrow (0.62), relief (0.14)
eyes	confusion (0.47), alarm (0.3)
like	relief (0.49), understanding (0.48)
forward	hope (0.85), understanding (0.74)
quickly	alarm (0.83), understanding (0.56)
lowers	sorrow (0.51), desire (0.32)

Table 3: Words and our model's highest corresponding predicted emotions.

Ablations We performed an ablation analysis on our model where we modified the descriptions of the robot that our model uses to train. We did this by removing or retaining stop words, nouns, verbs, adjectives, and adverbs. The part-of-speech tags of the words in the descriptions were chosen using TextBlob, and the model was trained as explained in 5.1. We use the same metrics (i.e., accuracy, precision, recall, f1 score) as Experiment 3.

The results are displayed in Table 4, demonstrating that the base model and the version trained with no stop words perform the best according to all metrics, as expected. The version of the model trained with only stop words performed second-worst given the F1 score, as we expect that no relevant emotional information is conveyed in stop words. The version of the model trained on only verbs had a slightly worse F1 score, indicating that verbs may not carry meaningful information for a model that is predicting perceived emotion. When comparing F1 scores of the versions of the model trained on only nouns or only verbs, it is clear that nouns are much more important to such a model than verbs, and this is further demonstrated by the high F1 score in the version of the model where verbs were removed. This shows that the composed descriptions represent complex language–the composed description is a better predictor of emotion.

	Acc	**Prec**	**Recall**	**F1**
Base	**90.4%**	**69.6%**	**31.4%**	**43.3%**
No Stop Words	89.9%	63.9%	30.3%	41.1%
Only Stop Words	87.8%	38.2%	7.4%	12.4%
Removed Nouns	88.8%	56.1%	18.3%	27.6%
Only Nouns	87.6%	41.2%	16.0%	23.0%
Removed Verbs	89.5%	64.4%	21.7%	32.5%
Only Verbs	87.5%	31.4%	6.3%	10.5%
Removed Adj	90.2%	68.5%	28.6%	40.3%
Only Adjectives	89.2%	67.6%	14.3%	23.6%
Removed Adverbs	90.3%	68.8%	30.3%	42.1%
Only Adverbs	87.9%	18.2%	1.1%	2.2%

Table 4: Results from ablation study, where descriptions of the robots's behaviors were modified before training.

Correlation with Sentiment We compared our results with a sentiment classifier to determine if simple sentiment classifiers capture nuances of emotion in different types of written text. We calculated sentiment scores with our model by first predicting the 16 emotions split into positive/negative valence pairs (as done in (McNeill and Kennington, 2019): *interest/alarm, understanding/confusion, relief/frustration, joy/sorrow, graditude/anger, hope/fear, surprise/boredom, desire/disgust*, then multiplied the probabilities of the 8 negative emotion labels by -1, then averaged the scores to obtain a sentiment score between -1 and 1, making our scores comparable to the sentiment scores given by the TextBlob sentiment classifier. We then calculated the correlation of determination between the sentiment scores from our model and TextBlob. The R^2 result was 0.11, indicating a poor correlation. We believe that our model learns robot-specific sentiment and context that a general model does not account for. We also note that the descriptions for this task represent an observer's interpretation of a behavior, whereas sentiment classification is tasked with the emotional content within a written text; this shows that applying sentiment classification may not help in certain contexts. This point is further reinforced by a high R^2 value of 0.84 between the sentiment scores calculated from the true values in the data and our model, and a R^2 value of -0.51 between the sentiment scores calculated from the data and the sentiment scores from TextBlob.[3]

5.2 Experiment 2: Classifying Emotion Labels from Robot Behaviors

In this experiment, we use the robot states represented as Novikova features to train and evaluate a classifier to predict what people perceive about the emotional display of Cozmo's behaviors. The purpose of this experiment is two-fold: (1) to compare with the results of Experiment 1 to ascertain if robot behaviors represented as internal state updates contain comparable information to descriptions of those same behaviors (i.e., the concreteness assumption compared against the abstractness assumption), and (2) to build a model of affect that a grounded semantic model can use to ground into for Experiment 3.

Task & Procedure Given features that represent a robot behavior (i.e., the Novikova features described above) predict one or more of the 16 emotion labels. We perform a 10-fold cross validation and average the results of each fold. No individual behavior was represented both in the training and test set for each fold.

Model & Training Because our feature vectors are small, and because we do not have much training data, our approach uses a multi-label K-Nearest Neighbor (KNN) classifier (Zhang and Zhou, 2007) to map from features to a distribution over class labels (tests using other classifiers, such as neural networks, resulted in worse performance, likely due to the nature of the features and small amount of training data).[4] The only parameter that was needed for the KNN classifier was number of neighbors, which we set to 5 to balance generalizability and performance in our task.

Metrics To compare directly to prior work and Experiment 1, we report two metrics to give an overall understanding of how our model performs this classification task in decreasing degrees of constraint:
average accuracy: Because any given behavior could have multiple labels, we take the prediction distribution and for every affect that received a probability of more than 0.5 in that distribution (0.5 was determined using the development set), we counted that as a positive guess for that affect. We compared this to the labels (i.e., by comparing two binary vectors) and compute the accuracy for each behavior, then take the average. This will seem inflated as many zeros in the binary vectors will increase the accuracy, but it allows our model to predict multiple labels, which better follows what people do when interpreting emotions (i.e., this is the same as micro averaging, as done in Experiment 1). ***F1 fscore***: We compute the F1 score using micro averaging. We also report precision, recall, and F1 score for each individual affect.

The baseline we are comparing against is the best resulting setting and ablation reported in McNeill and Kennington (2019) (termed M&K below). Note that their work used the raw features from the robot including facial, audio, and internal state features. Their model was a multi-layer perceptron.

5.3 Results

The results are shown in Table 5 and Figure 3. The former shows a direct comparison to M&K; we are reporting state-of-the-art results in both metrics despite only using internal state updates (their work included facial and audio features, as produced by the robot) and our classifier is much simpler. Moreover, the use of a KNN classifier suggests that there are vectors (at least for the Cozmo robot) which are prototypical for a particular affect, even if individual behaviors for each affect appear to be very different. This is further evidenced by the individual results in Figure 3, which show high overall for each individual emotion (with the exception of boredom). Comparing the results to Experiment 1, we find that, given a succinct representation like the Novikova features, mapping from internal states to emotions performs comparably to a mapping from words (i.e., in the descriptions) to emotion. We are therefore able to capture an equivalent amount of information by applying either the concreteness assumption or the abstractness assumption to this task. Taken together, the model from this experiment which maps from internal states (represented as Novikova features) to a distribution over 16 emotions could potentially be used as a simulated

[3] We opted to not compare with other fine-grained emotion models like Abdul-Mageed and Ungar (2017) because because the goal is to compare the model from Experiment 1 with the model from Experiment 3.

[4] Multi-label KNN uses Bayesian Inference to compute a probability for each label.

Approach	Assumption	Feature Set	Avg accuracy	Avg f1
M&K	Concrete	raw internal states	71.0%	55.0%
Exp. 1 (BERT)	Abstract	text description	90.4%	43.3%
Exp. 2 (KNN)	Concrete	Novikova features	**91.0%**	**57.0%**

Table 5: We predict emotion labels using various feature sets derived from either the abstractness or concreteness assumption. Results show that our KNN model using the Novikova features (exp 2) performs better than the McNeill and Kennington (2019) prior work on the same task. Results of using the concreteness assumption in Exp. 2 are comparable to the abstractness assumption using text description as the features set in Exp. 1.

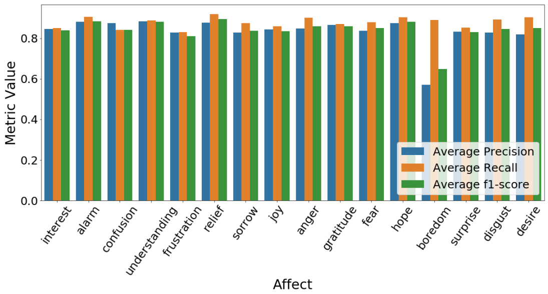

Figure 3: Average precision, recall, and F1 score for each emotion label for Experiment 2.

affective representation that a grounded classifier could ground into, which we explore in Experiment 3.

5.4 Experiment 3: Grounding Descriptions into Robot Behaviors and Affective States

In this experiment, we seek to answer the question: *Can a model for grounded lexical semantics ground into internal robot states as well as a representation of affect derived from those internal states?* We use the WAC model explained above as the model of grounded lexical semantics, the Novikova features as the robot state representation, and the trained model from Experiment 2 as a vector representation of affect (i.e., we treat it was an emotional substrate to ground into).

Task & Procedure Given the words in a description, we randomly select n distractor behaviors from the test set and apply the features of those distractors along with the gold behavior to the classifiers for each word in the description using the WAC model. Following Schlangen et al. (2016), the sum of the probabilities from these classifiers results in the semantic appropriateness score for each behavior. The model correctly identifies the gold behavior when that behavior is assigned the highest score.

We repeat this for 1-19 distractors d and report the accuracy for each d using a 10-fold cross-validation. Each behavior is represented twice in the dataset (once for each description), but we ensure that the same behavior never occurs in both train and test sets. We use words that are represented 5 or more times in the training data, resulting in a vocabulary of 347 words. Based on evaluations from one of the folds, we determined that for each positive instance of a word in the corpus, we should randomly choose 3 negative instances.

Model & Training Prior work using WAC has traditionally used logistic regression classifiers and multilayer perceptrons. Here, we take inspiration from Experiment 2 and use KNN for the WAC classifiers, as they are using the same set of features as the KNN classifier in Experiment 2 used. We trained an individual KNN for each word in our vocabulary with the nearest neighbor parameter set to 5 to balance generalizability and fitness.

A shortcoming of WAC is the independence assumption which fails to capture more complex semantic patterns within the text. However, the purpose of this experiment is not to capture the various semantic dependencies within the text, but to measure the comparative strength of our feature sets. WAC provides a simple and effective means by which to compare constructed affect to its lower-level source: robot state data.

We report results for when WAC only used the Novikova features directly (i.e., grounding into internal robot states represented as the 9 Novikova features), and also using the trained model from Experiment 2 as a representation of affect (16 features) for a total of 25 features.

Metrics The metric we report is accuracy of the model in choosing the correct retrieved behavior from the list of distractors, given the description. The baseline for this experiment are the the other two ablations: when WAC model performance without emotion labels, and the WAC model that only uses emotion labels.

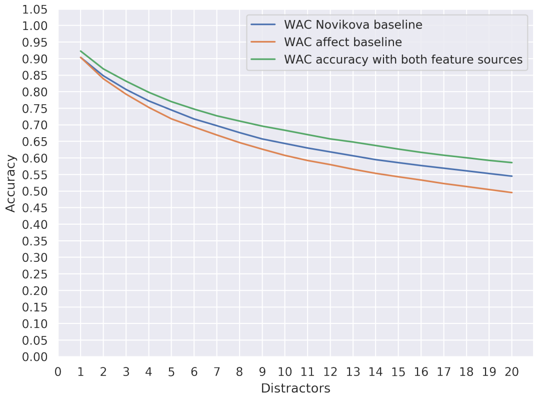

Figure 4: Experiment 3 Results: WAC applied to Novikova and affect feature types; accuracy for each ablation compared to number of distractors.

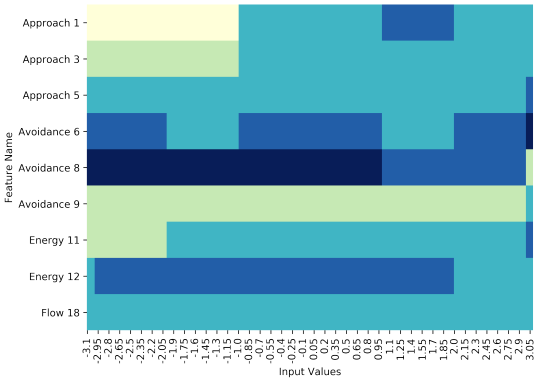

Figure 5: Heat map generated by ranging over all possible features and applying them to the WAC classifier for the verb *turns*; darker colors denote higher probabilities.

Results The results are shown in Figure 4. Overall, the model performs with high accuracy when there is only 1 distractor and we observe decreases in accuracy as the number of distractors increases. The WAC classifier performs well on its own given the 9 Novikova features alone, but it performs noticeably better when including our derived representation of affect from Experiment 2 (particularly when there are a larger number of distractors). Our results show that a model of grounded semantics can effectively ground into a predicted representation of affect. This is a very informative result: automated systems clearly don't have their own intrinsic biological / chemical affect, yet semantics can benefit from using a proxy of affect derived from human *perceptions*.

Analysis of Verbs In this section, we show how WAC learns verb semantics from the Novikova features by ranging over all possible values for each feature (i.e., we passed the Novikova features through a feature normalization transform, resulting in features ranging from -3.1 to 3.1), holding all other features to 0 when they are not being ranged over and using a WAC classifier to classify each possible combination of features.

Figure 5 illustrates what the *turns* WAC model learned by using a heatmap (darker colors denote higher probabilities returned by WAC). We can see from the figure that higher values in Avoidance 8 (i.e., wheel movement) is the most informative feature, where most values that denote some change in movement result in higher probabilities for the *turns* WAC classifier. From this we learn that a verb (used to denote a robot action, such as *turn*) will have the expected feature representation using the Novikova features in a KNN classifier. We noticed similar expected results from other verbs such as *tilts* (high sensitivity to Approach 3), *plays* (wide range of values for all features between Approach 3 and Flow 18) and *jerks* (high sensitivity to any feature relating to head and light movement with high energy).

6 Conclusion

This work represents a contribution to a growing shift towards human-centered, affective computing (Picard, 2000; Mohammad and Ovesdotter Alm, 2015). People who use natural language are also emotional beings; this has implications for the kinds of systems researchers develop that humans will interact with, as well as how the semantics of words will be acquired, represented, and applied in those systems; in particular, dialogue systems that are used for spoken interaction with multimodal agents, such as robots. This work helps shed light on how leveraging emotion and affect leads to more accurate semantic models.

Our work agrees with the claims in Lücking et al. (2019) which explains how distributional representations are not enough. The partial meaning of many words can be derived abstractly from nearby words (Firth, 1957), but words keep company with physical objects, entities, and situations as well as with other words. The complete semantic meaning of many words includes the world, perceived and represented through physical modalities such as affect.

For future work, we will expand our repertoire of robot emotional displays by producing novel emotional behaviors for different robot platforms. We will perform studies to determine what kinds of emotional behaviors make a robot more amenable to performing collaborative tasks with humans, specifically tasks where the robot must learn new words from human collaborators and what those words denote. Our work has implications for how humans interact with robots and the semantics of the words that robots learn from humans: the meaning of a word is a function of what that word refers to in a scope of lexical and physical context.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Ameeta Agrawal, Aijun An, and Manos Papagelis. 2018. Learning Emotion-enriched Word Representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 950–961, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hassan Alhuzali, Muhammad Abdul-Mageed, and Lyle Ungar. 2018. Enabling Deep Learning of Emotion With First-Person Seed Expressions. In *Proceedings of the Second Workshop on Computational Modeling of People{'}s Opinions, Personality, and Emotions in Social Media*, pages 25–35, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Muhannad Alomari, Paul Duckworth, David C Hogg, and Anthony G Cohn. 2017. Natural Language Acquisition and Grounding for Embodied Robotic Systems. In *In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

Lisa Feldman Barrett. 2017. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU : On Meaning , Form , and Understanding in the Age of Data. In *Association for Computational Linguistics*.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. *arXiv*.

Pasquale A Della Rosa, Eleonora Catricalà, Gabriella Vigliocco, and Stefano F Cappa. 2010. Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods*, 42(4):1042–1048.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

John R Firth. 1957. *Papers in Linguistics, 1934–1951*. Oxford University Press, Oxford, UK.

Ben Goertzel, Cassio Pennachin, Samir Araujo, Fabricio Silva, Murilo Queiroz, Ruiting Lian, Welter Silva, Michael Ross, Linas Vepstas, and Andre Senna. 2010. A general intelligence oriented architecture for embodied natural language processing. In *Artificial General Intelligence - Proceedings of the Third Conference on Artificial General Intelligence, AGI 2010*, pages 13–18.

Krystyna Grabski, Laurent Lamalle, and Marc Sato. 2012. Contrle prédictif et codage du but des actions oro-faciales. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1: JEP*, pages 289–296.

Mark Johnson. 2008. *The meaning of the body: Aesthetics of human understanding*. University of Chicago Press.

C. Kennington and D. Schlangen. 2015. Simple learning and compositional application of perceptually groundedword meanings for incremental reference resolution. In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, volume 1.

Douwe Kiela and Stephen Clark. 2015. Multi-and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal.

Richard D Lane and Lynn Nadel. 2002. *Cognitive Neuroscience of Emotion*. Oxford University Press.

Andy Lücking, Robin Cooper, Staffan Larsson, and Jonathan Ginzburg. 2019. Distribution is not enough: going Firther. In *Proceedings of the Sixth Workshop on Natural Language and Computer Science*, April, pages 1–10. Association for Computational Linguistics.

David McNeill and Casey Kennington. 2019. Predicting Human Interpretations of Affect and Valence in a Social Robot. In *Proceedings of Robotics: Science and Systems*, FreiburgimBreisgau, Germany.

Saif Mohammad and Cecilia Ovesdotter Alm. 2015. Computational Analysis of Affect and Emotion in Language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Lisbon, Portugal. Association for Computational Linguistics.

Daniele Moro and Casey Kennington. 2018. Multimodal Visual and Simulated Muscle Activations for Grounded Semantics of Hand-related Descriptions. In *Proceedings of the 22nd Workshop onthe Semantics and Pragmatics of Dialogue*.

Jekaterina Novikova, Gang Ren, and Leon Watts. 2015. It's Not the Way You Look, It's How You Move: Validating a General Scheme for Robot Affective Behaviour. In *Human-Computer Interaction – INTERACT 2015*, pages 239–258, Cham. Springer International Publishing.

Rosalind W Picard. 2000. *Affective computing*. MIT press.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

David Schlangen, Sina Zarriess, and Casey Kennington. 2016. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1213–1223.

Sichao Song and Seiji Yamada. 2018. Designing Expressive Lights and In-Situ Motions for Robots to Express Emotions. In *Proceedings of the 6th International Conference on Human-Agent Interaction - HAI '18*, pages 222–228, New York, New York, USA. ACM Press.

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. 2016. Learning Multi-Modal Grounded Linguistic Semantics by Playing " I Spy ". In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*.

Gabriella Vigliocco, Stavroula Thaleia Kousta, Pasquale Anthony Della Rosa, David P. Vinson, Marco Tettamanti, Joseph T. Devlin, and Stefano F. Cappa. 2014. The neural representation of abstract words: The role of emotion. *Cerebral Cortex*.

Peng Xu, Andrea Madotto, Chien-Sheng Wu, Ji Ho Park, and Pascale Fung. 2018. Emo2Vec: Learning Generalized Emotion Representation by Multi-task Training. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 292–298, Brussels, Belgium. Association for Computational Linguistics.

Min-Ling Zhang and Zhi-Hua Zhou. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*.