An Integrated Evaluation of Perception, Interpretation, and Narration

Nicole Maslan[1], Melissa Roemmele[2], and Andrew S. Gordon[2]

[1]Claremont McKenna College, 888 Columbia Ave, Claremont, CA 91711
[2]University of Southern California, 12015 Waterfront Drive, Los Angeles, CA 90094
nmaslan16@students.claremontmckenna.edu, roemmele@ict.usc.edu, gordon@ict.usc.edu

Abstract

In this paper, we describe our efforts to create an evaluation tool to aid in the development of artificial intelligence systems that integrate perception, reasoning, and language abilities. Based on an early and influential study by social psychologists Fritz Heider and Marianne Simmel, we created 100 short movies depicting the motions of two triangles and a circle around a box with a hinged opening. For each movie, we provide quantitative information about each object's trajectory, a formal description of the actions that can be perceived in each object's behavior, a formal interpretation of the social situation that is depicted, and a short English narration of the interpreted events.

Perception, Interpretation, and Narration

In an early and influential study, social psychologists Fritz Heider and Marianne Simmel (1944) presented subjects with a short animated film depicting the motions of two triangles and a circle around a box with a hinged opening. Asked to describe what they saw, subjects responded with coherent narratives of the social interactions among three characters, with interpretations of the motivations, intentions, and emotions of the anthropomorphized shapes. From an artificial intelligence perspective, these subjects engaged in a particularly interesting suite of cognitive tasks, effortless for people but challenging for today's technologies. What would it take to build an artificial intelligence system that could watch a movie in the style of Heider and Simmel, and generate a textual narrative indistinguishable from one written by a human?

We imagine that successfully building such a system would require the integration of three emerging technologies from three different areas of artificial intelligence research. First is the problem of *perception*. The system must recognize the sequence of actions intended by the animator in the trajectories of abstract shapes. In the original Heider-Simmel film, the larger of the two triangles jabs at the smaller one, pushing it up against the side of the box. The circle and the smaller triangle kiss and dance after being separated for some time. Procedurally, the task is one of segmentation and labeling of time-series data, and could be approached in much the same manner as seen in pen-based handwriting recognition systems and video-based gesture recognition systems.

Second is the problem of *interpretation*. The system must explain the observed actions in anthropomorphic terms, ascribing motivations, intentions, and emotions to each of the characters. In the original Heider-Simmel film, the circle cowers in fear as the larger triangle corners it inside the box. The larger triangle becomes furious when the two other character escape, and destroys the box in a fit of rage. Procedurally, we see this interpretation task as one of logical abduction, where a knowledgebase of human sociology and psychology is used to find a parsimonious explanation of the observed actions.

Third is the problem of *narration*. The system must generate a coherent discourse to communicate its interpretation of the perceived actions in a natural, humanlike style. By Heider and Simmel's report (1944), subjects authored narratives rich with mentalistic phrases: the girl hesitates, she doesn't want to be with the first man, the girl gets worried, is still weak from his efforts to open the door, they finally elude him and get away, he is blinded by rage and frustration. Procedurally, the task is that of discourse generation, where the relational structure of the interpretation informs the discourse structure, and the semantics of the interpretation inform the lexical choices.

An Integrated Evaluation

Individually, each of the tasks of perception, interpretation, and narration pose difficult artificial intelligence challenges, although none seem insurmountable given recent progress in their respective research areas. However, the integrated task presents new challenges and new questions. Recent work in segmentation and labeling of time-series data has focused on supervised approaches that output probability distributions. How are these distributions to be successfully integrated with the formalisms used in logical abduction? Moreover, how can the results of the interpretation process be used to adjust these probability distributions, such that the interpretation influences what is perceived? Recent work on discourse generation in computational linguistics has pursued a strongly statistical approach, realizing the most probable surface text given the underlying representation. How is this process changed when the underlying representations are structural interpretations, containing both informative and commonsense assertions?

To tackle both the individual and integrated research challenges of perception, interpretation, and narration, we developed a new evaluation tool based on the original Heider-Simmel film. This evaluation tool consists of 100 short movies in the style of Heider and Simmel. To support the development of an integrated processing pipeline, this collection includes gold-standard annotations for each of the intermediate processes needed to translate these movies into coherent textual narratives, described below.[1]

Trajectory Data

We created 100 movies in the style of Heider and Simmel using a custom authoring tool (Gordon & Roemmele, 2014), which allowed movies to be quickly created by dragging the shapes across the screen of a tablet computer. Each movie is 90 seconds or less, and is encoded as a time-series dataset of position and rotation values for each object (big triangle, little triangle, circle, and a door). An example is the movie that animates the following sequence:

> *A small triangle and big triangle are next to each other. A circle runs by and pushes the small triangle. The big triangle chases the circle.*

Observable Actions

We authored gold-standard annotations of the recognizable actions in each animation, consisting of 16 unique one-character actions (e.g. *dance*, *jump*) and 30 unique two-character actions (e.g. *fight*, *kiss*). For these labels, we drew from the set of actions used in the website *Triangle Charades* (Roemmele et al., 2014), which aims to crowdsource the collection of training data for automated action recognition. These annotations were authored as conjunctions of literals in first-order logic, as in the example below:

$$approach'(E1,C,LT) \land push'(E2,C,LT) \land chase'(E3,BT,C) \land seq(E1,E2,E3)$$

Choice of Interpretations

The interpretation of each of the 100 movies is subjective, but some interpretations are consistently favored over others by human observers. For each movie, we authored a pair of possible interpretations as English-language sentences, and had human raters select the more plausible. We then formalized each interpretation as a conjunction of first-order logic literals (Maslan et al., 2015).

Why does the big triangle chase the circle?

a. *The big triangle is angry that the circle pushed the small triangle, so it tries to catch the circle.*
   $angryAt'(e4,BT,C)$

b. *The big triangle and circle are friends. The big triangle wants to say hello to the circle.*
   $friend'(e5,BT,C) \land goal'(e6,e7,BT) \land greet'(e7,BT,C)$

English-language Narrations

Finally, we authored a short English-language narration of each film as an example output of a perception, interpretation, and narration pipeline. Systems that successfully execute these processes should produce textual narratives that cannot be consistently distinguished from those written by humans that watch the same films.

> *The circle is trying to get away from the cops and pushes the small triangle to get out of its way. The big triangle feels attacked that the circle pushed its friend and chases after the circle too.*

Acknowledgments

This research was supported by the Office of Naval Research, grant N00014-13-1-0286. This material is based upon work supported by the National Science Foundation under Grant No. 1263386.

References

Gordon, A. and Roemmele, M. (2014) An Authoring Tool for Movies in the Style of Heider and Simmel. In A. Mitchell et al. (Eds.): International Conference on Interactive Digital Storytelling, Singapore, November 3-6, 2014 (ICIDS-2014), LNCS 8832, pp. 49--60. Springer International Publishing Switzerland (2014).

Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. American Journal of Psychology, 13, 1944.

Maslan, N., Roemmele, M., and Gordon, A. (2015) One Hundred Challenge Problems for Logical Formalizations of Commonsense Psychology. Twelfth International Symposium on Logical Formalizations of Commonsense Reasoning, Stanford, CA, March 23-25, 2015.

Roemmele, M., Archer-McClellan, H., and Gordon, A. (2014) Triangle Charades: A Data-Collection Game for Recognizing Actions in Motion Trajectories. 2014 International Conference on Intelligent User Interfaces, February 24-27, 2014, Haifa, Israel.

[1] Available at: https://github.com/asgordon/TriangleCOPA