Abductive Explanation of Emotions

Paul O'Rorke*
Phone & Fax: (714) 854-2894
E-Mail: ororke@ics.uci.edu
Department of Information
and Computer Science
University of California, Irvine
Irvine, CA 92717

Andrew Ortony†
Phone: (708) 491-3500
E-Mail: ortony@ils.nwu.edu
Institute for the Learning Sciences
Northwestern University
1890 Maple Avenue
Evanston, IL 60201

Abstract

Emotions and cognition are inextricably intertwined. Feelings influence thoughts and actions which in turn give rise to new emotional reactions. We claim that people infer emotional states in others using common-sense psychological theories of the interactions between emotions, cognition, and action. We have developed a situation calculus theory of emotion elicitation representating knowledge underlying common-sense causal reasoning involving emotions. We show how the theory can be used to construct explanations of emotional states. The method for constructing explanations is based on the notion of abduction. This method has been implemented in a computer program called AMAL. The results of computational experiments using AMAL to construct explanations of examples based on cases taken from a diary study of emotions indicate that the abductive approach to explanatory reasoning about emotions offers significant advantages. We found that the majority of the diary study examples cannot be explained using deduction alone, but they can be explained by making abductive inferences. The inferences provide useful information relevant to emotional states.

Introduction

Explaining people's actions often requires reasoning about emotions. This is because experiences give rise to emotional states which in turn make some actions more likely than others. For example, if someone strikes another person, we may explain the aggression as being a result of anger. As well as reasoning about actions in terms of emotional states, we can reason about emotional states themselves. Explaining emotional states requires reasoning about the cognitive antecedents of emotions. In the right context, we might reason that a person was angry because he or she had been insulted. This paper focuses on explanations of this kind.

We present a computational model of the construction of explanations of emotions. The model is comprised of two main components. The first component is a situation calculus theory of emotion elicitation. The second component is a method for constructing explanations. The representation of emotion eliciting conditions is inspired by a theory of the cognitive structure of emotions proposed by Ortony, Clore, and Collins (1988). In addition to codifying a set of general rules of emotion elicitation inspired by this theory, we have also codified a large collection of cases based on diary study data. We have implemented a computer program that constructs explanations of emotions arising in these scenarios. The program constructs explanations based on a first order logical abduction method.

Abductive Explanation

Peirce used the term abduction as a name for a particular form of explanatory hypothesis generation (Peirce, 1931–1958). His description was basically:

The surprising fact C is observed;
But if A were true,
C would be a matter of course,
hence there is reason to suspect that A is true.

Since Peirce's original formulation, many variants of this form of reasoning have come to be known as abduction. Examples of abduction methods proposed in AI research include abductive approaches to diagnosis (Peng & Reggia, 1990) and natural language comprehension (Hobbs, Stickel, Martin, & Edwards, 1988). We focus on a logical view of abduction advocated by Poole (e.g., Poole, Goebel, & Aleliunas, 1987). In this approach, observations O are explained given some background knowledge expressed as a logical theory T by find-

*Supported in part by National Science Foundation Grant Number IRI-8813048.

†Supported in part by National Science Foundation Grant Number IRI-8812699.

Table 1: Elicitation Conditions for 20 Emotion Types

$$
\begin{aligned}
joy(P,F,S) &\leftarrow wants(P,F,S) \wedge holds(F,S).\\
distress(P,F,S) &\leftarrow wants(P,\overline{F},S) \wedge holds(F,S).\\
happy\_for(P_1,P_2,F,S) &\leftarrow joy(P_1,joy(P_2,F,S_0),S).\\
sorry\_for(P_1,P_2,F,S) &\leftarrow distress(P_1,distress(P_2,F,S_0),S).\\
resents(P_1,P_2,F,S) &\leftarrow distress(P_1,joy(P_2,F,S_0),S).\\
gloats(P_1,P_2,F,S) &\leftarrow joy(P_1,distress(P_2,F,S_0),S).\\
hopes(P,F,S) &\leftarrow wants(P,F,S) \wedge anticipates(P,F,S).\\
fears(P,F,S) &\leftarrow wants(P,\overline{F},S) \wedge anticipates(P,F,S).\\
satisfied(P,F,S) &\leftarrow precedes(S_0,S) \wedge hopes(P,F,S_0) \wedge holds(F,S).\\
fears\_confirmed(P,F,S) &\leftarrow precedes(S_0,S) \wedge fears(P,F,S_0) \wedge holds(F,S).\\
relieved(P,\overline{F},S) &\leftarrow precedes(S_0,S) \wedge fears(P,F,S_0) \wedge holds(\overline{F},S).\\
disappointed(P,\overline{F},S) &\leftarrow precedes(S_0,S) \wedge hopes(P,F,S_0) \wedge holds(\overline{F},S).\\
proud(P,A,S) &\leftarrow agent(A,P) \wedge holds(did(A),S) \wedge praiseworthy(A).\\
self\_reproach(P,A,S) &\leftarrow agent(A,P) \wedge holds(did(A),S) \wedge blameworthy(A).\\
admire(P_1,P_2,A,S) &\leftarrow agent(A,P_2) \wedge holds(did(A),S) \wedge praiseworthy(A).\\
reproach(P_1,P_2,A,S) &\leftarrow agent(A,P_2) \wedge holds(did(A),S) \wedge blameworthy(A).\\
grateful(P_1,P_2,A,S_1) &\leftarrow agent(A,P_2) \wedge holds(did(A),S_1) \wedge precedes(S_0,S_1) \wedge\\
&\quad causes(A,F,S_0) \wedge praiseworthy(A) \wedge wants(P_1,F,S_1) \wedge holds(F,S_1).\\
angry\_at(P_1,P_2,A,S_1) &\leftarrow agent(A,P_2) \wedge holds(did(A),S_1) \wedge precedes(S_0,S_1) \wedge\\
&\quad causes(A,F,S_0) \wedge blameworthy(A) \wedge wants(P_1,\overline{F},S_1) \wedge holds(F,S_1).\\
gratified(P,A,S_1) &\leftarrow agent(A,P_2) \wedge holds(did(A),S_1) \wedge precedes(S_0,S_1) \wedge\\
&\quad causes(A,F,S_0) \wedge wants(P,F,S_1) \wedge holds(F,S_1) \wedge praiseworthy(A).\\
remorseful(P,A,S_1) &\leftarrow agent(A,P_2) \wedge holds(did(A),S_1) \wedge precedes(S_0,S_1) \wedge\\
&\quad causes(A,F,S_0) \wedge wants(P,\overline{F},S_1) \wedge holds(F,S_1) \wedge blameworthy(A).
\end{aligned}
$$

ing some hypotheses H such that

$$H \wedge T \vdash O.$$

In other words, if the hypotheses are assumed, the observation follows by way of general laws and other facts given in the background knowledge.

We construct explanations using an abduction engine based on an early approach to mechanizing abduction described in (Pople, 1973). The method is implemented in a PROLOG meta-interpreter called AMAL. It takes as input a collection of PROLOG clauses encoding theories. One theory represents background knowledge, another captures the facts of the case at hand. An observation to be explained is given as a query. AMAL is also given an operationality criterion and an assumability criterion. The output includes an explanation of the given observation, possibly including some assumptions that must be made in order to complete the explanation.

In general, many explanations are possible and it is important to constrain the search to avoid large numbers of implausible hypotheses and explanations. In early experiments, we found that the abduction engine conjectured large numbers of implausible causal relationships. This problem was solved by disallowing assumptions of the following forms:

$$preconditions(A,F)$$
$$causes(A,F,S)$$

In other words, the abduction engine was not allowed to assume that an arbitrary fluent might be a precondition for an action, nor was it allowed to assume unprovable cause-effect relationships between actions and fluents.

Emotion Elicitation

Our first order logical theory of emotion elicitation contains rules covering eliciting conditions of twenty emotion types (see Table 1). In addition, we have coded variants of a number of them, details of which have been omitted due to space constraints. (See O'Rorke & Ortony, 1992 for a presentation of

the full theory.)

The theory draws upon knowledge representation work on situation calculus (McCarthy, 1968) and conceptual dependency (Schank, 1972). It includes axioms that support causal reasoning about actions and other events that can lead to emotional reactions.

For example, the first law below mediates positive and negative effects of actions. The second law states that a precondition of a physical transfer from one location to another is that one must first be at the initial location. The remaining laws state the effects of a physical transfer.

$$holds(F, do(A, S)) \leftarrow causes(A, F, S) \wedge poss(A, S).$$
$$poss(ptrans(P, To, From, T), S) \leftarrow holds(at(T, From), S).$$
$$causes(ptrans(P, To, From, T), \underline{at(T, To)}, S)$$
$$causes(ptrans(P, To, From, T), \overline{at(T, From)}, S).$$

Emotion types are represented as fluents and their eliciting conditions are encoded in rules. As examples, consider the rules for the emotion types *fear* and *relief*, shown in Table 1. The *fear* rule captures the idea that people may experience fear if they want an anticipated fluent not to hold. Relief may be experienced when the negation of a feared fluent holds. Fear usually occurs before the fluent holds. Note that, although many examples of fear involve expectations, we use the predicate *anticipates* in an effort to suggest the notion of "entertaining the prospect of" a state of affairs. The purpose of this is to avoid suggesting that hoped-for and feared events necessarily have a high subjective probability.

Explaining Emotions

In this section, we use an example to illustrate the abductive construction of explanations involving emotions. The example is based on data taken from a diary study of emotions. Most of the subjects who participated in the study were sophomores at the University of Illinois at Champaign-Urbana. They were asked to describe emotional experiences that occurred within the previous 24 hours. They typed answers to a computerized questionaire containing questions about which emotion they felt, the event giving rise to the emotion, the people involved, the goals affected, and so on. Over 1000 descriptions of emotion episodes were collected, compiled, and recorded on magnetic media. We have encoded over 100 of these examples using our situation calculus representation language. The following case provides examples of *relief* and *fear*.

Mary wanted to go to sleep.
Karen returned.
T.C. finally left her place.
Mary was relieved.

The case is encoded as shown in Table 2. The case fact says that Mary wants sleep. The query asks why Mary is relieved that T.C. is not at her home in the situation that results after T.C.'s departure. T.C.'s departure occurred in the situation resulting from Karen's return. (Note the abbreviations for the relevant situations at the bottom of the Table.)

The explanation shown in Table 2 was constructed automatically by AMAL. The program works by backward chaining on observations to be explained. It tries to reduce the observation to known facts by invoking general laws (e.g., causal laws of situation calculus and laws of emotion elicitation). In this case, the eliciting condition for relief is invoked in order to explain Mary's relief. This generates new questions that must be answered, and so on. The resulting explanation (shown in Table 2) states that Mary is relieved that T.C. is no longer at her home. The explanation assumes that Mary fears T.C.'s presence in her home because she wants T.C. not to be in her home but she anticipates that he will be there. A deeper explanation connecting this desire and anticipation to Mary's desire for restful sleep should be possible. For example, the presence of T.C. might interfere with Mary's sleep. The explanation of his abcence does not include the possibility that he may have been driven away by Karen's return. But it does serve to illustrate the use of causal laws to infer negative fluents relevant to emotional reactions. In this case, since T.C. moved from Mary's home to another location, it can be inferred that he is no longer at Mary's home.

Discussion

Like the example of relief and fear, the majority of the cases in the diary study data require assumptions. The kinds of assumptions needed include missing preconditions, goals, prospects, and judgements. In the example, the assumption that T.C. was at Mary's home in the initial situation helped explain why he was there after Karen came home. This in turn was a precondition for T.C.'s leaving Mary's home. The example also required an assumption that Mary wanted T.C. to go somewhere else in order to explain Mary's fear that T.C. would be at her home. Assumptions about other's goals also occur in explaining emotions that involve the "fortunes of others." Abductive assumptions about other mental states include assumptions about whether agents anticipate events. In the example of *relief*, it was necessary to assume that Mary anticipated T.C.'s continued (unwelcome) presence in her home. Assumptions about judgements of blameworthiness and praiseworthiness are important in explaining a number of emotions not present in the example.

The explanation constructed in the example, and

Table 2: Explanations of Relief and Fear

```
Case Facts
    wants(mary, sleep(mary), _)
Query
    why(relieved(mary, not at(tc, home(mary)), s2))
Explanation
    relieved(mary, not at(tc, home(mary)), s2)
        precedes(s1, s2)
        fears(mary, at(tc, home(mary)), s1)
            [wants(mary, not at(tc, home(mary)), s1)]
            [anticipates(mary, at(tc, home(mary)), s1)]
        holds(not at(tc, home(mary)), s2)
            causes(ptrans(tc, _29887, home(mary), tc), not at(tc, home(mary)), s1)
            poss(ptrans(tc, _29887, home(mary), tc), s1)
                holds(at(tc, home(mary)), s1)
                    not causes(ptrans(karen, home(mary)), not at(tc, home(mary)), s0)
                    [holds(at(tc, home(mary)), s0)]
                    poss(ptrans(karen, home(mary)), s0)
Abbreviations
    s1=do(ptrans(karen, home(mary)), s0)
    s2=do(ptrans(tc, _591, home(mary), tc), s1)
```

many other explanations (see O'Rorke & Ortony, 1992), could not have been constructed by the abduction engine without its abductive inference capability, given the background knowledge and codifications of the cases provided with the observations to be explained. Given the same information, a purely deductive PROLOG-style interpreter would fail to find an explanation. Admittedly, the knowledge base could conceivably be extended so that some assumptions could be eliminated and replaced by deductive inferences. For example, if knowledge of ethics and standards of behavior could be provided, the number of assumptions in explanations requiring judgements of blameworthiness and praiseworthiness could be reduced. But it is not likely that all relevant preconditions, desires, prospects, and judgements can be provided in advance.

Related and future work

We give a complete description of the situation calculus of emotion elicitation in (O'Rorke & Ortony, 1992). That paper also contains additional examples and details of the mechanism used to generate explanations.

A previous study formalizing commonsense reasoning about emotions is summarized in (Sanders, 1989). This work takes a deductive approach, using a deontic logic of emotions. The logic focuses on a cluster of emotions involving evaluations of actions — including what we have called admiration, reproach, remorse, and anger. The evaluation of actions is ethical, and involves reasoning about obligation, prohibition, and permission. The logic was used to solve problems involving actions associated with ownership and possession of property (e.g., giving, lending, buying, and stealing) by proving theorems. For example, the fact that Jack will be angry was proved given that he went to the supermarket, parked his car in a legal parking place, and when he came out, it was gone. It is not clear whether the theorems were proved automatically or by hand so questions of complexity of inference and control of search in the deontic logic remain unanswered. We have argued that abduction offers advantages over deduction alone when applied to the task of constructing explanations involving emotions. And our situation calculus of emotion elicitation is more comprehensive than the deontic logic for emotions in that it covers more emotion types. But our approach could benefit from Sanders' treatment of ethical evaluations. We hope to undertake a detailed comparison and integration of the best parts of the two approaches in future work.

The present work focuses on explaining emotions in terms of eliciting situations. But while situations give rise to emotional reactions, emotions in turn give rise to goals and actions that change the state of the world. Applications such as plan recognition

will require a theory specifying causal connections between emotions and subsequent actions. For a brief description of a system for recognizing plans involving emotions, see Cain, O'Rorke, and Ortony (1989). This paper also describes how explanation-based learning techniques can be used to learn to recognize such plans. For a fuller discussion of reasoning about emotion-induced actions, see Elliott and Ortony (1992).

In Ortony, Clore, and Foss (1987) about 270 English words are identified as referring to genuine emotions from an initial pool of 600 words that frequently appear in the emotion research literature. In another study, 130 of these emotion words were distributed among 22 emotion types. Some emotion words map to several different types, e.g., "upset" is compatible with distress, anger, or shame. Many words map to the same type. Encoding the relationship between the affective lexicon and the emotion types is an important topic for future research aimed at automatically processing natural language text involving emotions.

Conclusion

We have developed a theory of the cognitive antecedents of emotions and an abductive method for explaining emotional states. We sketched a computer program, an abduction engine implemented in a program called AMAL, that uses the theory of emotion elicitation to construct explanations of emotions. We presented an explanation of an example based on a case taken from a diary study of emotions.

The most important advantage of our approach to explanatory reasoning about emotions is that abduction allows us to construct explanations by generating hypotheses that fill gaps in the knowledge associated with cases where deduction fails. In most cases, emotional states cannot be explained deductively because they do not follow logically from the given facts. The abduction engine explains the emotions involved in these cases by making assumptions including valuable inferences about mental states such as desires, expectations, and the emotions of others.

Acknowledgments

We thank the reviewers for suggestions that improved the paper. Terry Turner kindly provided diary study data. Steven Morris, Tim Cain, Tony Wieser, David Aha, Patrick Murphy, Stephanie Sage, Clark Elliott, and Milton Epstein participated in various stages of this work.

References

Cain, T., O'Rorke, P., & Ortony, A. (1989). Learning to recognize plans involving affect. In A. M. Segre (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 209–211). Ithaca, NY: Morgan Kaufmann.

Elliott, C., & Ortony, A. (1992). *Point of view: Modeling the emotions of others.* Manuscript submitted for publication.

Hobbs, J. R., Stickel, M., Martin, P., & Edwards, D. (1988). Interpretation as abduction. *Proceedings of the Twenty Sixth Annual Meeting of the Association for Computational Linguistics* (pp. 95–103). Buffalo, NY: The Association for Computational Linguistics.

McCarthy, J. (1968). Programs with common sense. In M. Minsky (Ed.), *Semantic Information Processing* (pp. 403–418). Cambridge, MA: MIT Press.

O'Rorke, P., & Ortony, A. (1992). *Explaining emotions* (Technical Report 92-22). Submitted for publication. Irvine: University of California, Department of Information and Computer Science.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions.* New York: Cambridge University Press.

Ortony, A., Clore, G. L., & Foss, M. A. (1987). The referential structure of the affective lexicon. *Cognitive Science,11*(3), 361–384.

Peirce, C. S. S. (1931–1958). *Collected papers of Charles Sanders Peirce (1839–1914).* Cambridge, MA: Harvard University Press.

Peng, Y., & Reggia, J. A. (1990). *Abductive inference models for diagnostic problem solving.* New York: Springer-Verlag.

Poole, D. L., Goebel, R., & Aleliunas, R. (1987). Theorist: A logical reasoning system for defaults and diagnosis. In N. Cercone, & G. McCalla (Eds.), *The Knowledge Frontier: Essays in the Representation of Knowledge.* New York: Springer-Verlag.

Pople, H. E. (1973). On the mechanization of abductive logic. *Proceedings of the Third International Joint Conference on Artificial Intelligence* (pp. 147–152). Stanford, CA: Morgan Kaufmann.

Sanders, K. E. (1989). A logic for emotions: a basis for reasoning about commonsense psychological knowledge. In E. Smith (Ed.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 357–363). Ann Arbor, MI: Lawrence Erlbaum and Associates.

Schank, R. C. (1972). Conceptual dependency: A theory of natural language-understanding. *Cognitive Psychology,3*(4), 552–631.