Inside Case-Based Explanation

Roger C Schank, Alex Kass, Christopher K Riesbeck (eds)
Published: Lawrence Erlbaum Associates, 1994

Mike Brown and George Paliouras
Dept. of Computer Science
The University of Manchester

Case-Based Reasoning (CBR) is a rapidly growing field in AI. It is a memory-driven approach to problem solving; completed solutions to problems are stored in an 'episodic' memory, retrieved and adaptively applied to solve new problems. "Inside Case-based Reasoning" [4], the popular predecessor to this book, is a well written introductory text for the subject.

This book is an attempt to expand the theory of CBR from a problem solving technique to a model of cognition. The main philosophical claim is that Case-Based Explanation (CBE) underlies understanding and learning (p21):

> "To understand is to satisfy some basic desire to make sense of what one is processing, to learn from what has been processed and formulate new desires about what one wants to learn. Our question is how people do this and how machines might do this."

The book has some of the practical, 'hands-on' nature of its predecessor. The fundamental theoretical ideas are centred on the essential components (i.e. case retrieval, explanation evaluation and adaptation) of a CBE system. The book is divided into three parts; the first is a relatively brief review of the proposed model of explanation, the second is a collection of six detailed descriptions of individual systems and the third is an annotated listing of a simplified prototype for the proposed CBE system.

Part I of the book is a condensation of an earlier book by Schank; "Explanation Patterns: Understanding Mechanically and Creatively" [5], which is a much more comprehensible and entertaining text for the first-time reader. The crux of the theory is, however, adequately described here.

The discussion starts by broadening the current scope of explanation within the field of AI. The much maligned Turing Test is rejected in favour of a test that gauges the level of understanding of a reasoning agent in terms of its ability to generate explanations; not only for others but also for itself. A persuasive argument is given that the main way an agent learns is through the creative generation of explanations of perceived anomalies, where an anomaly is defined as a mismatch between a new situation and the knowledge stored in memory.

The interesting questions are "What is an explanation?" and "How are explanations formed?". A content-theory answer is adopted for the first question. It is postulated that memory is populated by a collection of *Explanation Patterns (XPs)*. XPs are script-like structures capturing stereotypical causal descriptions constructed from primitives such as; *agents*, *goals*, *beliefs* and *actions*. The level of generality of an XP varies from the characterisation of a single anomalous event (e.g. explaining the death of Janis Joplin) to descriptive structures applicable across one or more domains (e.g. being murdered for an inheritance).

Explanation Questions (EQ) are proposed as the basis by which anomalies within a new situation can be detected and applicable XPs to resolve these anomalies retrieved. Hence, XPs are the 'cases' to be adaptively applied to each new problem (i.e. anomaly). So, the answer to the second question is that explanations are not *formed* as such but *reused*. Adaptation is seen as the root of creativity as it allows the mutation of pre-existing XPs when used in novel, anomalous situations.

Part II of the book (chapters 3 to 9) is an attempt to make the theory of the first part more concrete. Chapter 3 is a brief overview of a historically important explanation system, **SWALE**, which is central to most of the book. As such, chapter 3 is largely just a reiteration of the model of explanation as one of retrieving, instantiating and adapting XPs. In the remainder of the second part of the book, individual systems relating to specific parts of **SWALE** are described in detail.

In chapters 4 and 5, the issue of retrieving previous explanations from memory is considered. Chapter 4 addresses a fundamental issue of case retrieval, namely, the problem of determining the appropriate abstraction level for the memory indices. Indices that are highly abstract can be an extremely powerful guide for retrieving relevant cases but are expensive to accurately ascribe to an input case. Conversely, the specific information that is readily available in an input case is a poor guide for retrieval, particularly when cases from a variety of domains are stored. The proposal in this chapter is that the appropriate level of abstraction for indices for explanation of planning anomalies is that of the generic functional criteria typically captured in proverbial advice (c.f. [3]). The proposal is that such indices can be derived by a dynamic, incremental retrieval strategy that combines bottom-up access of candidate knowledge structures in memory with the top-down inference of relevant indices from a pool of currently active knowledge structures (c.f. [2]). This approach is embodied in the **Anon** system, whose ability to solve the problem of retrieving explanations is discussed in detail.

Chapter 5 focusses on the indexing vocabulary required for explanation retrieval. A 5-stage model for describing events involving agents is put forward covering: *theme*, *goal*, *plan*, *action* and *effect*. The model is used in the **Abby** system to represent sociological stories. Through numerous examples, it is shown that the 5-stage model can be used to represent these stories in an abstract framework, allowing indexing and other reasoning (e.g. detection of goal conflict) to

be performed. However, the reasoning is limited to a simple form of pattern matching. The underlying principle behind this chapter (and much of the book) is a commitment to extensive knowledge engineering for the construction of a domain-specific explanation system.

Chapter 6 mainly covers the crucial issue of evaluating a candidate explanation retrieved for an anomalous situation; the system described is **Accepter**. The work is interesting in two respects. Firstly, a context-sensitive model of evaluation is given, where an explanation's worth is determined taking into account the current goals of the reasoner (p192):

> "By retrieving explanations indexed under the same anomaly type as the current problem, a case-based explainer can efficiently generate candidate explanations that are relevant to its needs for information."

Secondly, an elegant recursive model for anomaly detection and resolution is provided. Anomaly detection in explanation is achieved by matching the explanation against stereotypes, for example, restricting the type of agent involved in a particular action. An explanation used to resolve a specific anomaly can give rise to further anomalies.

In chapter 7, the **AQUA** system is described. This is claimed to be an alternative view of understanding and explanation built on a basic question asking and answering mechanism. In reality, the model is very similar to those of the other systems described in the book. Questions in the **AQUA** model are memory structures specifying both a knowledge requirement and the purpose for which that knowledge is to be used once it is made available. In this respect, questions are classified in terms of different types of "knowledge goal" that an understanding agent inherently possesses (such as the need to detect and resolve anomalies).

There are several appealing aspects to the explanation model embodied in **AQUA**. As for **Accepter**, the model of explanation is recursive and relativistic; question-answering can spawn further questions which are determined by the knowledge structures currently held by the reasoner. An interesting classification of why a system may fail to explain is also provided, covering three general causes of failure: encountering a *novel situation*, having an *incorrect world model* and *misindexing* of knowledge stored in memory. It is also pointed out that the process of question-answering is a means for focussing the inference that is performed by an understanding agent. However, although it is recognised that the main computational problems involve the searches in question-generation, answering and evaluation, no efficient mechanisms for these processes are put forward. The operation of **AQUA**, as for other systems is illustrated by detailed examples.

Chapters 8 and 9 cover the final aspect of CBE, namely adaptation. Chapter 8 describes **TWEAKER**, which modifies an explanation by the selection and application of retrieval strategies stored in memory. In keeping with the rest

of the book, the chapter focusses on the types of adaptation strategy that exist and ignores the issue of how an appropriate strategy is selected. A similar problem to that encountered in **Anon** also exists here; highly general (syntactic) adaptation strategies are too weak to be reliably applied, but more specific adaptation strategies tend to be too limited in the scope of their applicability. It is recognised that the appropriate level of abstraction of adaptation strategies is as meta-knowledge, capturing different types of search strategy that can be carried out within the memory, in order to replace the anomalous features of an explanation. The strategies are categorised as *generalise*, *substitute* or *specialise*. These strategies are designed to cover (respectively) the three general causes of explanation-failure identified in the **AQUA** system.

The final chapter in part II of the book describes the **Brainstormer** system. **Brainstormer** is concerned with the adaptation of plans rather than explanations. The particular problem addressed is how a system that possesses a memory of highly general planning advice (such as is depicted in proverbs) can apply such advice to specific planning problems. The proposed solution is to redescribe the specific problem in terms of the abstract plan vocabulary via the process of *lambda abstraction*. This is achieved through the encoding of structures within memory that represent possible resolution strategies; essentially rules that represent ways of recognising instances of a given abstract concept. Again, there is recursion in the basic mechanism of **Brainstormer**, as the results of applying a redescription structure may be intermediate concepts that themselves need to be redescribed in order to fully comply with a candidate abstract planning concept. One concern with the **Brainstormer** system is in how well it scales up. The advice offered by alternative proverbs is often contradictory (c.f. "Too many cooks spoil the broth" and "Many hands make light work" [3]). Resolving such ambiguity requires detailed knowledge of the specific planning situation. However, the redescription structures in **Brainstormer** are described in more abstract terms and it is questionable whether they possess the inferential power to selectively guide redescription for systems involving a larger memory of abstract plans.

Part III of the book presents the code (in Common Lisp) for a cut-down version of the **SWALE** system. The components discussed in the preceding sections, including an explanation retriever, accepter and tweaker, are covered by this example implementation. Also provided are the frame-like structures used to represent the various types of knowledge required for the proposed CBE system, along with examples of each knowledge structure. The overall structure of the program, as well as the details of the functionality of each module, are well described.

The provision of this annotated code is arguably the best aspect of this publication as it provides some backing to the abstract ideas discussed in the theory, as well as forming a good basis for anyone wanting to build an explaining program. The only worry is that the code is, by necessity, extremely simple which means that the more interesting and sophisticated reasoning processes of the full-blown

SWALE system are not adequately demonstrated.

In summary, the book provides a very informative overview of CBE, a plausible model of the explanation process, based on the creative reuse of old explanation experiences. As with any of Schank's publications, the text is thought provoking and the general principles behind the model of explanation (and indeed understanding *per se*) are highly appealing. The book briefly presents the theoretical underpinning of the work and describes in detail systems which implement the model.

The philosophy adopted in this book is that of a 'content theory'; the power of an explanation system is seen to derive directly from the types of knowledge structure and their organisation in memory. On that basis, the book does not provide a rigorous treatment of either the theory, in terms of formally defining the concepts involved, or the systems, by means of analysing the properties of the algorithms that are used. Instead, the authors tend to rely on the use of extended examples, protracted and *ad hoc* categorisations (of explanation patterns, anomaly types, etc.) and intuitive justifications to support many of the strong claims made within the book. Any attempt to establish necessary and sufficient properties of what generally constitutes an explanation (let alone what constitutes a *good* explanation) is largely avoided and those seeking a more rigorous treatment of explanation should look elsewhere (e.g. [1]). Moreover, much of the proposed model of CBE requires tackling potentially massive search spaces and complex decision making. The issue of computational intractability is treated dismissively (p33):

> "..., as in the old joke about the woman who agrees to sex for $1 million and then is offended with a suggestion that she do it for $5, we are still just arguing about price... Determining how much to search and what to consider when is an interesting topic, but not one of great theoretical interest."

The book does provide some interesting comparisons to related approaches. For example, creative reuse of old knowledge structures is argued for as a more efficient mechanism than traditional inference chaining for complex reasoning (p76). The use of abduction for providing explanations from memory is compared to the deductive process employed by traditional Explanation Based Learning (EBL) techniques (p172). CBE is preferred in this respect because it provides the ability to choose between a set of possible explanations on general structural criteria. In terms of knowledge usage, CBE is based on large, rich knowledge structures whereas EBL relies on chains of simple inference rules (p223). This comparison gives a clear picture of the fundamental differences between the proposed case-based model and other approaches to explanation, which is beneficial to the reader unfamiliar with CBE, but who has some experience with alternative methods. More generally, the book largely refers to standard AI technology to

describe the various systems, hence the text is comprehensible for those with a grounding in AI and lacking experience in the area of CBR.

Unfortunately, the book is poorly edited; as the bulk of the theory is presented in independently written chapters, describing different but related systems, this leads to unacceptable repetition, often obscuring the major issues. As a result, the book lacks the clarity to make it an ideal undergraduate text. For the experienced researcher already familiar with Schank's work, this book provides little in the way of new material. The book is perhaps best suited to those commencing research in the area of cognition.

References

[1] Peter Achinstein. *The Nature Of Explanation.* Oxford University Press, 1983.

[2] Charles Eugene Martin. *Direct Memory Access Parsing.* PhD thesis, Yale University, 1991.

[3] Christopher Owens. Domain-Independent Prototype Cases for Planning. In Janet Kolodner, editor, *Proceedings of the Case-Based Reasoning Workshop*, Florida, May 1988. Morgan Kaufmann Publishers.

[4] C.K. Riesbeck and R.C. Schank. *Inside Case-Based Reasoning.* Lawrence Erlbaum Associates, 1989.

[5] Roger C Schank. *Explanation Patterns. Understanding Mechanically and Creatively.* Lawrence Erlbaum Associates, 1986.