Learning from Textbook Knowledge: A Case Study

William W. Cohen
Computer Science Department
Rutgers University
New Brunswick, NJ 08903
wcohen@cs.rutgers.edu

Abstract

One of the "grand challenges for machine learning" is the problem of learning from textbooks. This paper addresses the problem of learning from texts including omissions and inconsistencies that are clarified by illustrative examples. To avoid problems in natural language understanding, we consider a simplification of this problem in which the text has been manually translated into a logical theory. This learning problem is solvable by a technique that we call *analogical abductive explanation based learning (ANA-EBL)*. Formal evidence and experimental results in the domain of contract bridge show that the learning technique is both efficient and effective.

Introduction

One of the "grand challenges for machine learning" that Dietterich suggested in his address at the 1989 Machine Learning Conference is the problem of learning from textbooks. A great deal of knowledge from a huge variety of domains has been codified in textbook form; however, this knowledge is not directly usable by computer programs.

Some of the problems involved in learning from textbooks are natural language understanding problems. However, these are not the only problems, and perhaps not even the major problems. The information in even a well-written textbook contains inconsistencies and omissions that would make it useless if transcribed directly into logic. Learning from textbook knowledge is thus an important and challenging machine learning problem.

The text on contract bridge used in this experiment [Sheinwold, 1964] illustrates some of the specific technical problems involved in learning from textbook knowledge. The first three chapters of the book were manually translated into Horn clause logic, and found to contain several undefined terms, and many pieces of advice that are contradictory. However, the popularity of this text (which has sold millions of copies, and is still in print after 23 years) shows that these logical shortcomings are not problematic to human readers.

Many of the inconsistencies and ambiguities in the text are clarified by *illustrative examples*. For instance, in discussing rules for bidding two-suited hands [Sheinwold, 1964, page 16], Sheinwold says "The general rule is, if your suits are unequal in length, bid the longer one; if your suits are equal in length, bid the higher one."[1] but immediately adds that "You have to disregard this general rule on some hands". After a short digression into what defines a "biddable suit", he presents fifteen examples that clarify this general rule. Such passages make it clear that the bidding rules presented are over-general, and that understanding the examples is necessary to fully understand the concept being taught. This should not be surprising — common sense tells us that understanding examples and working exercises is a crucial part of learning from textbooks.

To summarize, the problem addressed in this paper is *learning by understanding examples from a textbook*. Unlike previous work (*e.g.*, [VanLehn, 1987]), we assume that the textbook knowledge is imperfect, and that examples are used to clarify flaws in the knowledge; the problem of learning from textbook knowledge will thus be treated as a a special instance of the more general problem of *learning from understanding examples using an approximate theory.*

Description of the problem

In this case study, the first three chapters of a text on the game of contract bridge [Sheinwold, 1964] were manually translated into Horn clause logic. These three chapters taught opening bids, a small but reasonably interesting part of the game of bridge; Sheinwold devotes 34 pages to this subject.

Most of the rules required to understand the examples of the first three chapters are clearly and explicitly presented, which made transcription into logic straightforward. However, the direct transcription into

[1] Terminology from the bridge domain will be used in this paper only in presentation of examples. Hence familiarity with contract bridge, while helpful to the reader, is not strictly necessary.

logic resulted in a theory that had several flaws.

- Certain key properties (for example, whether a hand had "good length in the major suits") were not defined.
- Several of the rules in the theory were over-general, and clearly not intended to be used in all the situations in which they were applicable.

Augmenting the theory with commonsense knowledge partially solved the first problem. For each undefined property, it was possible to restrict the number of possible definitions to a small number of reasonable candidates. For example, the definition of "good length in the majors suits" could be inferred to be "having length in the major suits greater than L" for some unknown minimal length L. Completing the theory could thus be reduced to the problem of choosing the correct definition of each undefined term.

Apart from the incompleteness of the theory, the over-general rules also caused the theory to be inaccurate, as measured by a sample test also taken from [Sheinwold, 1964]. Because of the over-general rules, *no* completion could achieve a score of better than 12 out of a possible 16 problems: on the remaining 4 problems, the theory suggested multiple bids, some of which were incorrect.[2]

We concluded that: *in order to complete and correct the theory, additional information was needed.* The obvious source of additional information is the examples that accompany the text. How can the information in these examples be extracted?

One possibility is to incorporate into the theory *all possible completions* of the undefined predicates. The result is a theory that is complete but over-general. The problem of incorporating the examples now becomes a *theory specialization problem* [Flann and Dietterich, 1989]:

Given: 1) an over-general theory T_i,
2) a set of examples of correct uses of T_i,
Find: a theory T_s that specializes T_i, and that only leads to correct predictions.

The specialization T_s of the initial theory T_i will be called the *target theory.* In learning from textbooks, T_s is the concept being taught (in this case, opening bids) expressed as a set of Horn clauses, and T_i includes all of the definitions of terms (such as "major suit"), all possible definitions of undefined terms (such as "length in majors") and whatever rules are given in the textbook (such as the bidding rules given in [Sheinwold, 1964]).

In the context of learning from textbooks, the theory specialization problem ideally should be solved *without using any knowledge from outside the textbook* (*e.g.*, generalization hierarchies, etc) other than commonsense knowledge. Showing that this constraint is satisfied in a learning system that is not purely automatic can be difficult. In our experiments, for example, one insidious source of additional knowledge is the representational choices made in transcribing the text into logic. Efforts were made to make these choices in a consistent and natural manner, and also to avoid introducing additional knowledge not explicitly present in the textbook; the latter policy was followed even at the cost of omitting some information that could (perhaps) be inferred by an intelligent reader. Nevertheless, some representational choices that affect learning needed to be made; thus the skeptical reader may wish to view this work as learning from a synthetic theory that is believed (by the authors) to be prototypical of the sort of theories that could be automatically derived from a textbook. See [Cohen, 1989] for a detailed description of the transcription process.

The following problems make theory specialization difficult in this domain.

1. The initial theory T_i can produce multiple inconsistent explanations of an example; *i.e.*, it suffers from the *multiple explanation problem* [Rajamoney and DeJong, 1988].
2. The target theory T_s is disjunctive; *i.e.*, no single rule is sufficient to describe all correct opening bids.
3. The initial theory T_i and target theory T_s are relational, not propositional.

Problems 1 and 2 rule out use of the mEBG and IOE techniques discussed in [Flann and Dietterich, 1989]; the fact that negative examples are also present also argues against the appropriateness of these techniques (with each hand, Sheinwold presents a list of correct bids; possible bids not on the list are thus by inference negative examples). Problem 2, and the lack of a generalization language, rules out use of the technique of incremental version-space merging [Hirsh, 1989]. Problem 3 rules out use of the MIRO algorithm [Drastal *et al.*, 1989], which could also be considered a theory specialization technique. Finally, the presence of an almost-correct initial theory suggests that traditional inductive learning techniques, which cannot use this information directly, are not appropriate.

The techniques that seem most appropriate to this problem are the techniques described in [Pazzani, 1988] and [Fawcett, 1989] for using explanation-based learning (EBL) on a theory that generates multiple inconsistent explanations. These researchers have identified heuristics for choosing between multiple inconsistent explanations. If heuristics could be found that make the correct choices, then a refinement theory T_s could be formed by simply disjoining the results of performing explanation based generalization (EBG) on the chosen explanations.

In investigating such approaches, however, a final obstacle was uncovered:

[2] It is possible for a hand to have more than one correct bid in Sheinwold's bidding system.

4. No single fixed level of operationality is appropriate to learning.

Introduction of an operationality predicate is somewhat problematic in any case: since no operationality predicate is explicitly given in the textbook, using any operationality predicate at all violates the principle of using no knowledge from outside the textbook. However, in this context, an operationality predicate represents knowledge about what features are relevant to the bidding problem; this knowledge could reasonably be inferred by a reader. For instance, the bidding rules given by Sheinwold are presented at a fairly high level; rules are typically given in terms of high level features of hands such as the number of high card points, the length of biddable suits, etc. It seems reasonable for a reader to assume that lower-level features are *not* relevant to bidding rules. This and similar arguments can be used to justify a choice of an operationality predicate.[3]

However, the sample test from [Sheinwold, 1964] indicates that this level of operationality is too low for a standard explanation based learner to achieve good performance. In two of the sixteen test problems, the correct bid is supported by an explanation that was different (at the chosen level of operationality) from *every* explanation of *every* training example. A consequence of this is that *no rule generated by performing EBG on some explanation of a training example would apply to these test cases.*

For both of these anomalous test case, although the correct explanation is not *identical* to the explanation of any of the training examples, the correct explanation is very *similar* to the explanation of some training example; for example, the explanation of the correct bid of 1 club for the test case ♠ KJ642 ♡ A5 ♢ 3 ♣ AQ732 differs in only one subproof from the explanation used to justify the bid of 1 club on the training example ♠ KQJ75 ♡ 5 ♢ 62 ♣ AJ963. This particular problem could be handled by marking the predicate *opening_strength* as operational. Unfortunately, if this were done, then in some other cases, every rule learnable from a training example would be over-general.

Considerations such as these suggest that some sort of simple analogical reasoning strategy is needed, where bids can be accepted or rejected based on consideration of training examples with similar, but non-identical explanations. This in turn suggests that explanation-based analogical reasoning techniques such as those described in [Huhns and Acosta, 1987; Kedar-Cabelli, 1987] could be combined with explanation-selection heuristics to solve the learning problem. These analogical reasoning techniques use EBG with a artificially high level of operationality to produce rules that match any potential analogies. These rules are over-general, in the sense that not all instances that match these rules should be treated the same as the training example from which the rule was formed. One can view such a rule as an *explicit* representation of the generalizations that would be *implicitly* made by an analogical reasoner.

In this research, a slightly different approach was taken to analogy. For each training example, instead of generating a single very general rule, a large number of somewhat general rules were produced, each corresponding to a class of analogical instances. What one would like to do is to pick general rules that *only* match instances that should be treated the same as the training example. A key insight is that *choosing the right generalization can be done with the same techniques used to solve the multiple explanation problem.*

The learning algorithm

The learning algorithm used is shown in Figure 1. It takes as input a set of positive examples S^+, a set of negative examples S^-, a theory T, an operationality predicate $\mathcal{O}$, and an additional parameter k, which will be discussed shortly. The algorithm is called *analogical abductive explanation based learning* (ANA-EBL), since the theory can generate multiple inconsistent explanations and hence is similar in character to the abductive theories described in [O'Rorke, 1988; Pazzani, 1988]. The reader is referred to [Mitchell *et al.*, 1986] for definitions of terms such as "explanation structure", and for an algorithm for explanation based generalization.

The basic idea of the algorithm is simple. First, all possible generalizations of the positive training examples are enumerated, where "all possible generalizations" includes generalizations formed by first marking up to k internal nodes of some explanation structure for the example as "operational". These extended generalizations can be matched by a new problem with an explanation that differs from the training example in up to k subproofs. The parameter k is thus a constraint on how *similar* an new example must be to a training example in order to be treated analogously. Inconsistent generalizations (those that match some negative example) are then filtered out, and finally, a greedy set cover algorithm is used to find a minimal-sized disjunction of the remaining candidates which covers all the positive examples.

Unfortunately, space limitations make presentation of a detailed example of the algorithm impossible; the interested reader is referred to [Cohen, 1989]. An appendix to this paper containing an example and a short summary of some relevant formal results is also available from the author on request.

Note that with $k = 0$, no analogical reasoning takes place; in this case ANA-EBL simply uses the set covering and size heuristics to choose between multiple explanations and find the set of EBG rules that best describes the data.

[3] Again, the interested reader is referred to [Cohen, 1989] for a more detailed description of our choice.

Algorithm ANA-EBL($S^+, S^-, T, \mathcal{O}, k$):

1. Compute the explanation structure of every proof of every example in S^+.
2. For each explanation structure found in step 1, find the set of candidate rules that can be formed by
   (a) marking up to k internal nodes of the explanation structure as "operational"
   (b) applying the final stage of EBG to the resulting explanation structure.
3. Filter the set of candidate rules by removing any rule that covers an element of S^-.
4. Use a greedy set cover to find a small set of rules that accounts for all of the training examples:
   (a) Initially, let COV be the empty set.
   (b) Add to COV that candidate rule R that maximizes the ratio of the number of as-yet-uncovered examples explained by R to the size[a] of R.
   (c) Repeat step 4b until all examples have been covered.
5. Return the disjunction of the rules in COV.

[a] The size of R is defined to be the number of nodes in the explanation structure from which R was formed.

Figure 1: The ANA-EBL Learning Algorithm

It can be easily shown that this algorithm runs in time polynomial in the total size of the set of proofs for elements of S^+, but exponential in k. Of course, the number of proofs can be very large or even infinite; ANA-EBL is only efficient when this is not the case.

It can also be shown that ANA-EBL satisfies Valiant's criterion of efficient learnability [Valiant, 1984].

Theorem 1 (From [Cohen, 1989]) *Let n be the minimal size (over all samples) of any set of rules generated by the procedure above that correctly define the target theory T_s, and let $|T_i|$ be the number of Horn clauses in the initial theory T_i. Then with probability at least $1-\delta$, ANA-EBL will return a specialized theory that will have error (with respect to the probability distribution function D) less than ϵ if it is given only $m(\frac{1}{\epsilon}, \frac{1}{\delta}, n)$ examples chosen stochastically according to D, where*

$$m(\frac{1}{\epsilon}, \frac{1}{\delta}, n) = O(\max(\frac{1}{\epsilon}\log\frac{1}{\delta}, \frac{n\log|T_i|}{\epsilon}(\log\frac{n\log|T_i|}{\epsilon})^2))$$

Furthermore, there exists initial theories T_i such that every *algorithm that with probability at least $1-\delta$ returns a specialized theory that has an error of less than ϵ will require* at least *$m(\frac{1}{\epsilon}, \frac{1}{\delta}, n)$ examples, where*

$$m(\frac{1}{\epsilon}, \frac{1}{\delta}, n) = \Omega(\frac{1}{\epsilon}\ln\frac{1}{\delta} + \frac{n}{\epsilon})$$

Assuming that n is large and $\log|T_i|$ is small, the upper and lower bounds on m can be simplified to $O(\frac{n}{\epsilon}(\log\frac{n}{\epsilon})^2)$ and $\Omega(\frac{n}{\epsilon})$ respectively. So in learning from random examples, ANA-EBL is within a factor of $(\log\frac{n}{\epsilon})^2$ from the optimum.

Note also that as k increases, n (the optimal size of T_s) will decrease; the theory thus predicts that increasing k will improve the convergence rate of ANA-EBL.

Of course, in the context of learning from textbooks, training examples will be carefully chosen, not randomly selected; this is in fact one of the principle advantages of learning from textbooks. However, this result is still important; it suggests that ANA-EBL is robust with respect to which training examples are chosen, and reinforces the claim that ANA-EBL is a general learning algorithm — in particular, that it is not specific to the contract bridge domain.

Experimental results

To further evaluate the ANA-EBL algorithm, a series of experiments were conducted. The first set of experiments used as an initial theory T_i a transcription of the first three chapters of [Sheinwold, 1964], and used as training data the forty-eight examples used by Sheinwold in the first three chapters, with no additions, and five omissions. The omitted examples dealt with opening in third- or fourth-hand position; they were omitted because eliminating position information greatly simplified our representation of the bidding problem. As test data we used the relevant portion of the sample test in [Sheinwold, 1964]; this consisted of sixteen bidding problems.

The examples and test data taken from [Sheinwold, 1964] are a fair test of the learner in two ways; they are representatives of a naturally-occurring concept of some complexity, and they were chosen without knowledge of the learning algorithm. It is not an ideal test, however, because of its small size. To circumvent this problem, a program was written that randomly generated bridge hands and then opened them using hand-coded bidding rules. The hand-coded rules are a reasonable implementation of the bidding system presented in [Sheinwold, 1964]; for instance, they bid all of the problem hands in the sample test correctly. By using this program as a classifier, unlimited amount of training and test data can be generated; however, this introduces another source of potential bias, because the data no longer consists of true representatives of Sheinwold's *opening_bid* concept, but of representatives of our own interpretation of that concept.

Experiments with textbook data

The ANA-EBL algorithm was used to learn a bridge bidding strategy in two phases. First, the textbook theory was completed by adding all possible completions (subject to common-sense constraints) of the undefined predicates. The predicate *opening_strength* was then learned from the completed theory and the training examples, using ANA-EBL with $k = 0$. All of

Theory	Accuracy
Initial theory T_i	12/16
T_i with learned *opening_strength*	12/16
output of ANA-EBL ($k = 0$)	14/16
output of ANA-EBL ($k = 1$)	15/16
output of ANA-EBL ($k = 2$)	15/16

Table 1: Learning *opening_bid* from textbook data

the undefined predicates appear in the definition of *opening_strength*, and none of them appear anywhere else; hence learning this predicate is equivalent to simultaneously learning definitions of all of the undefined predicates. *Opening_strength* is the least general predicate that has this property and that is also "observable", in the sense that one can look at a training example and readily determine whether the hand in question makes *opening_strength* true.

The learned definition of *opening_strength* was then spliced into the original initial theory, and the predicate *opening_bid* was learned from the training examples, using ANA-EBL with $k = 0$ and $k = 1$. This predicate says which bids are correct opening bids for a given hand; teaching this is the main object of the first three chapters of the text. The learned predicates were then tested on the sample test. A problem was judged to be correct if no incorrect bids were suggested, and at least one correct bid was suggested. The results are shown in Table 1; for comparison, we also give the performance of the initial domain theory T_i, and of T_i with the learned version of the *opening_strength* predicate.

It would be preferable to learn *opening_bid* directly, rather than first learning *opening_strength* in a separate pass. However, because there are many possible definitions for each of several undefined concepts, the original theory generates hundreds of alternative explanations of each *opening_strength* goal. Our implementation can handle this degree of ambiguity, but only for extremely low values of k, so learning in a single pass would preclude experimentation with larger values of k. This issue is discussed more completely in [Cohen, 1989].

Experiments with random data

The experiments above show that the learning algorithm works well for a small set of carefully selected illustrative examples. Experiments were also performed with randomly selected data. The goals of these experiments were first, to measure the benefit of using "textbook cases" for examples, rather than randomly chosen examples; and second, to further test the hypothesis that analogical reasoning is necessary in this domain, but that limited analogical reasoning is sufficient to solve most problems. Some evidence for the latter hypothesis is given by the performance of the

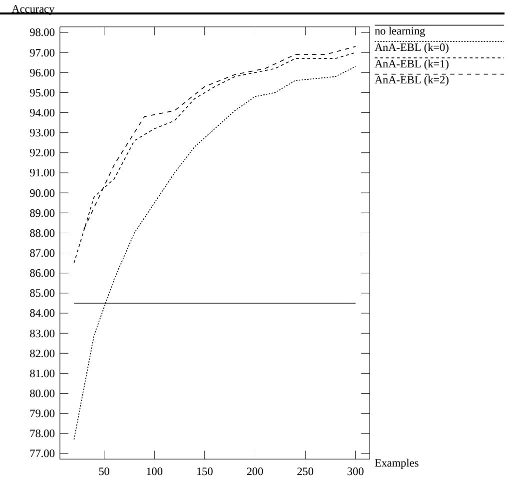

Figure 2: Learning *opening_bid* from random data

learning algorithm on the sample test, since $k = 1$ is sufficient to solve all but one of the test problems; however, because the sample test is so small, additional evidence is desirable.

Experiments were done in learning *opening_bid* from the initial theory with a correct definition of *opening_strength*. A test set of 1000 hands was generated and classified by the generation program. Then a separate training set of 300 hands was generated and classified by the program. ANA-EBL, with k = 0,1, and 2, was then given progressively larger subsets of the training set, and the accuracy of each T_s was measured by using it to classify the hands in the test set, and comparing the classifications to the correct ones. This experiment was repeated 10 times and the error rates were averaged, using the same test set in each trial. The result is the "learning curve" shown in Figure 2 that plots the accuracy of the hypothesis against the number of training examples.

The experiments show that the algorithm has good convergence properties, even on randomly selected data. In learning *opening_bid*, ANA-EBL with $k = 1$ learns substantially faster than ANA-EBL with $k = 0$, and only marginally slower than ANA-EBL with $k = 2$. This substantiates the conjecture that some analogical reasoning seems to be necessary, but that limited analogical reasoning is sufficient, and confirms the prediction made by the formal analysis that increasing k improves the rate of learning. It also shows that Sheinwold's examples are much more informative than randomly selected ones: about 7 times as many examples are needed to achieve comparable levels of accuracy using random examples.

Conclusions

To summarize, we have identified a problem that arises in learning from the knowledge in textbooks: the problem of learning from knowledge including *omissions and inconsistencies* that are clarified by *illustrative examples*. This learning problem is solvable by a technique that we call *analogical abductive explanation based learning (ANA-EBL)*. ANA-EBL actually solves the more general problem of *learning from understanding examples using an approximate theory*. It is an explanation based learning technique that combines techniques used to choose between multiple inconsistent explanations with explanation based analogical reasoning techniques along the lines of [Huhns and Acosta, 1987; Kedar-Cabelli, 1987].

Learning from textbook knowledge is a hard problem; even if the natural language problems are finessed by manually translating a theory into logic, the problem of correcting and completing the resulting theory is difficult. The major contribution of this paper is isolation of some subproblems involved in the general problem of learning from textbooks, and presentation of techniques that address these subproblems.

The techniques developed to solve this problem, however, are of independent interest. ANA-EBL is a *theory specialization* technique, like the techniques described in [Drastal *et al.*, 1989; Flann and Dietterich, 1989; Hirsh, 1989]; however, ANA-EBL works even in situations in which the original theory is relational and/or generates multiple inconsistent explanations, and in which the target theory is disjunctive. Experimental results indicate that the technique is effective on randomly selected examples, as well as on well-chosen "textbook" examples. ANA-EBL also builds on techniques described in [Pazzani, 1988] and [Fawcett, 1989] for solving the *multiple explanation problem* but extends these results by first, incorporating analogical reasoning techniques similar in flavor to those used in [Huhns and Acosta, 1987; Kedar-Cabelli, 1987]; and second, giving a precise way of weighting the complexity of an explanation and the number of observations that it covers, and justifying this heuristic with a formal analysis.

ANA-EBL is far from a complete solution to the problem of learning from textbook knowledge; however, we feel that it addresses at least some of the issues that must be confronted. An important topic for further research would be to integrate ANA-EBL with an automatic text understanding system. This would address the major shortcoming of our evaluation of ANA-EBL: in the process of manually translating a text into Horn clause logic, representational choices that affect learning inevitably must be made. Because of this, the skeptical reader may wish to view this work as learning from a synthetic theory that is prototypical of the sort of theories that could be automatically derived from a textbook.

Acknowledgements

I am grateful to Alex Borgida, Haym Hirsh, Chun Liew and many other members of the Rutgers AI community. The author is supported by an AT&T Fellowship. Initial stages of this research were supported by a Marion Johnson Fellowship.

References

(Cohen, 1989) William W. Cohen. Abductive explanation based learning: A solution to the multiple explanation problem. Technical Report ML-TR-26, Rutgers University, 1989.

(Drastal *et al.*, 1989) George Drastal, Gabor Czako, and Stan Raatz. Induction in abstraction spaces: A form of constructive induction. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, Michigan, 1989. Morgan Kaufmann.

(Fawcett, 1989) Tom Fawcett. Learning from plausible explanations. In *Proceedings of the Sixth International Workshop on Machine Learning*, Ithaca, New York, 1989. Morgan Kaufmann.

(Flann and Dietterich, 1989) Nicholas Flann and Thomas Dietterich. A study of explanation-based methods for inductive learning. *Machine Learning*, 4(2), 1989.

(Hirsh, 1989) Haym Hirsh. Incremental version space merging: A general framework for concept learning. PhD Thesis, Stanford University Department of Computer Science, 1989.

(Huhns and Acosta, 1987) Michael Huhns and Ramon D. Acosta. ARGO: An analogical reasoning system for solving design problems. Technical Report AI/CAD-092-87, MCC, 1987.

(Kedar-Cabelli, 1987) Smadar Kedar-Cabelli. Formulating concepts according to purpose. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, Seattle, Washington, 1987. Morgan Kaufmann.

(Mitchell *et al.*, 1986) T. Mitchell, R. Keller, and S. Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine Learning*, 1(1), 1986.

(O'Rorke, 1988) Paul O'Rorke. Automated abduction and machine learning. In *Proceedings of the 1988 Spring Symposium on Explanation Based Learning*, Pao Alto, California, 1988. AAAI.

(Pazzani, 1988) Michael Pazzani. Selecting the best explanation in explanation-based learning. In *Proceedings of the 1988 Spring Symposium on Explanation Based Learning*, Pao Alto, California, 1988. AAAI.

(Rajamoney and DeJong, 1988) Shankar Rajamoney and Gerald DeJong. Active explanation reduction: An approach to the multiple explanation problem. In *Proceedings of the Fifth International Conference on Machine Learning*, Ann Arbor, Michigan, 1988. Morgan Kaufmann.

(Sheinwold, 1964) Alfred Sheinwold. *5 Weeks to Winning Bridge*. Simon & Schuster, 1964.

(Valiant, 1984) L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11), November 1984.

(VanLehn, 1987) K. VanLehn. Learning one subprocedure per lesson. *Artificial Intelligence*, 31:1–40, 1987.