Note* 1103 Alan Bundy
July 4, 1996

Experimental Research in Automatic Theorem Proving

1 Introduction

The issue of experimental work in automatic theorem proving has recently arisen in a number of contexts. For instance, as an editor of the Journal of Automated Reasoning I was approached by a referee for assistance in assessing the merits of a paper reporting experimental work. As a result of this I formulated some assessment criteria and circulated them to the other JAR editors for discussion. Also I was approached by John McCarthy, who is concerned about the scientific importance of experimental work in AI, in general, and is commissioning a series of notes on experimental work in each area of AI. He asked me to write something about experimental work in ATP. This note records my thoughts on this issue. It also builds on BBN 863, which discussed the experimental methodology of our own research project.

2 The Importance of Experimental Work

When I presented my assessment criteria to the JAR editors, the reaction of Bob Boyer was effectively to deny the role of experimental work in ATP. He said:

> 'As for Alan's criteria, let me simply raise a sceptical eyebrow. Algorithmic complexity, as in theoretical computer science, is the solidly established scientific method for comparing the performance of algorithms. If someone these days tried to publish a paper that said that some sorting algorithm was better than some other sorting algorithm based upon a few empirical results, it's not very likely that he or she would get the paper published in any reputable scientific journal. The way to prove that some algorithm is better than another is to prove it, not to test them on a few examples. Once we leave the mathematical realm of the analysis of algorithms, we get into a swamp hardly more scientific than sociology, and an attempt to establish norms risks becoming tedious and political. "Evidence" is for lawyers and juries, proof is for mathematicians.'

And in later email he elaborated on this:

> 'However, even one of Gödel's early papers gave an example of how mathematical proof can be used to compare complexity of logical systems, showing that some logics only yield proofs much longer than proofs of certain theorems than do other logics.
>
> 'Similarly, in the comparison of provers, one ought to strive for a mathematical demonstration that one prover always finds proofs more quickly than another system, even if neither prover has a complexity that can be expressed in a convenient form, such as one can for sorting algorithms. Comparing two systems on a couple of examples has its informal uses, and for this reason such papers have value, but a comparison based upon a couple of examples cannot be regarded in any way as a scientific proof that one system is better than another, and attempts to establish criteria to evaluate such an unscientific business may be futile.'

Given Boyer's own experimental contributions, I was surprised by these views, which seem to me to question the importance of a lot of ATP research. I also think he is wrong, and for several reasons.

Firstly, the theoretical tools at our disposal are not up to the job. Consider, for instance, algorithmic complexity. Most mathematical theories are semi-decidable. So the amount of computation required to discover a proof is unbounded, *i.e.* is infinite as a function of the size of the theorem. This puts theorem proving beyond the pale of conventional complexity analysis. Alternatively, we could consider the size of the search space as a function of the length of proof. Most provers are super-exponential, because the branching rate increases with proof depth, *e.g.* as both the number and size of clauses increases. The amount of this growth is very theory dependent, making it difficult to make global statements about the complexity of a particular prover. Worst case performance will come with the most explosive theories and will be very bad indeed. Average case performance is hard to define since it is

*Notes in this series are for ϵ-baked ideas, for $1 \geq \epsilon \geq 0$. Only exceptionally should they be cited or distributed outwith the Mathematical Reasoning Group.

not clear what constitutes an average theory. Instead of algorithmic complexity we could appeal to the arithmetic hierarchy from mathematical logic. But most theorem provers lie at the same level in this hierarchy, so it fails to distinguish them.

Secondly, what theoretical results we do have tend to measure the less critical things. For instance, Boyer mentions Gödel's results on lengths of proofs. Unfortunately, length of proof is not the crucial statistic — it is size of search space we are interested in, and that is much harder to determine. There are some results on the elimination of redundancies in resolution proofs by refining the resolution rule, *i.e.* by imposing logical restrictions. However, there are no similar results available for the effect of heuristic restrictions. Yet logical restrictions have a much less dramatic effect on efficiency that heuristic ones.

In general, the theoretical tools we have are too crude. They fail to distinguish provers that we want to compare and where they do provide results they are often not the results we are most interested in. Of course, theoretical results are desirable and we should seek them and use them when they are available. But, for the moment and in the foreseeable future, they are not enough.

My third objection to Boyer's position is more fundamental. It is that certain kind of questions can *only* be answered empirically. My favourite example is whether Horn clause provers are useful. If a theory consists only of Horn clauses then all sorts of efficiency improving restrictions can be imposed on a prover, *e.g.* no ancestor resolution in SL-resolution, the restriction to unit resolution. However, Horn clause theories are of measure zero in the set of all theories, so on purely theoretical grounds you would expect these restrictions to be hardly ever applicable. In practice, Horn clauses are ubiquitous, and provers designed especially for them are a very worthwhile investment. The ubiquity of Horn clause theories is impossible to predict on theoretical grounds; only an empirical study could discover this. Similar, remarks apply to the investment in decision procedure and the frequency with which different kinds of decidable subproblems occur in theorem proving.

More generally, the efficacy of heuristic search control techniques can only be determined experimentally. For any given heuristic, one can always design problems which will defeat it[1]. The question is whether the heuristic will *usually* work. For this we need a notion of *typical* problem in the intended application. This is an inherently empirical notion. On theoretical grounds there is no reason to regard one problem as more typical than another, *i.e.* no reason to regard the pathological case which defeats our heuristic as atypical. We can only decide this by experimental testing on the example problems which are known to arise in practice.

3 Criteria for Assessing Experimental Work

In this §I present an updated version of the criteria I circulated to the JAR editors.

My starting position is that any piece of scientific research should state a hypothesis and then present evidence to support or reject it. We are not used to stating hypotheses explicitly in ATP papers, but with a bit of imagination we can see implicit hypotheses in most ATP work. Here are some examples of hypotheses which would require experimental supporting evidence.:

- *This new ATP technique gives improved performance*, where 'improved' might mean faster, less search, higher success rate, more natural proofs, friendly interaction, *etc.* The comparison would ideally be with a prover lacking this new technique, but otherwise identical.
- *This alternative ATP technique gives improved performance.* The possible meanings of 'improved' are the same, but the comparison now would ideally be with a prover which used a different technique for the same purpose, but was otherwise identical.
- *This prover can prove non-trivial theorems.* The non-trivial theorems might be well-known challenge problems, open conjectures, very big applications, *etc.* Hypotheses like these are a bit vague and need refining, *i.e.* what accounts for the success of the prover?

The onus is on the author of a hypothesis to identify the kind of evidence which would serve to support or refute it. Usually the prover will be tested on some example problems and then some analysis will be applied to the results. If the claim is about faster proof, higher success rate or smaller search spaces then the analysis might consist of the calculation and presentation of some statistics. If the claim is about more natural proofs or user-friendly interaction then the analysis might be a discussion of some case studies. In all these cases the analysis has to argue for two difficult to establish conclusions:

1. that the results extend beyond the current case studies; and
2. that the explanation of the results is the stated hypothesis and is not due to some other cause.

We elaborate on these issues below.

[1] One could adopt that as a defining characteristic of *heuristic* technique.

4 Establishing the Generality of Results

Theorem provers are applicable to an infinite number of widely varying problems. They can only be tested on a finite number of problems. We usually want to draw general conclusions from these experimental results. We, therefore, need to argue that the finite test set is representative of the infinite set of problems the prover was designed for. I addressed this issue in §4 of BBN 863 and cannot improve on the criteria I listed there, so I merely reproduce them below.

Obtaining representative sets of examples for testing a technique is a standard problem in science — and there are some standard solutions, *i.e.*:

- *Use a lot of dissimilar examples.* Showing that a technique is successful and/or efficient on a lot of dissimilar examples in itself provides evidence that the technique is widely applicable. In addition, it becomes harder to bias the example set in favour of the technique if it is large and heterogeneous.
- *Collect examples from an independent source.* If some third party has provided the examples and they have not been subsequently filtered then we have evidence that they are not biased in favour of any of the technique(s) being compared. However, care must be taken to show that the source is truly independent, *i.e.* is not, due to its origin, inherently biased. For instance, if the technique is aimed at some final application then it is good if the examples have arisen from that application. This might even be taken as a definition of *representative*.
- *Use the shared examples of the field.* If the field has developed a standard test set then it aids comparison of rival techniques if they are tested on this shared set. This does not make the test set representative. Indeed, there is a real danger that the test set is provided by the founders of the field and is biased in favour of the techniques that they developed. One of the aims of the field should be to modify the standard test set to make it more representative.
- *Use challenging examples.* It is easy to get a high success rate on easy problems, but this does not aid comparison with other techniques. The technique needs to be tested on examples that have caused problems for other techniques.
- *Use pathological examples.* Even when they do not have a shared set of test examples, most fields pass around a selection of pathological examples. These are of most use during the development phase for exposing over-simplifications, missing cases, false assumptions, *etc.* in the technique. Success on them also demonstrates the generality of the technique, although pathological examples are in some sense the opposite of representative[2].
- *Distinguish training from test examples.* The generality of a technique can be demonstrated by selecting a set of test examples and not looking at them or running the technique on them until it has been fully debugged on a separate training set. This guards against the accusation of special tuning. Of course, to be fair to the technique the training set should itself be representative of the range and complexity of problems that will be encountered in the test set.

5 Establishing the Explanation for the Results

The purpose of controlled experimentation is to simplify the task of explaining the experimental results. For instance, if we compare two provers which differ in only one respect then any difference in performance between them must be due to this sole difference. On the other hand, if the two provers differ in two respects then a difference in performance might be due to either difference in isolation or a combination of the differences. Unfortunately, most ATP experimentation is not well controlled. Compared provers usually differ in many respects. This makes allocation of credit very difficult.

This lack of experimental control is sometimes unavoidable. A difference in one respect might entail necessary difference in other respects. For instance, a natural deduction prover must necessarily use both different formulae and different inference rules from a resolution prover. Sometimes the lack of experimental control is avoidable in theory, but understandable in practice. For instance, a researcher might have invested heavily in time and money in software and hardware that is different from those used by a rival prover. Unfortunately, the lack of experimental control is also often due to sloppy experimental design and is wholly avoidable. Such poor science should be penalised by referees and other critics.

Fortunately, we can often circumvent lack of experimental control by more careful analysis of the experimental results. For instance, suppose two provers are being compared on speed. We can investigate the slower prover

[2] For this reason I have renamed these examples 'acute'. The intuition behind the name is that they present some key difficulty in a sharp and unencumbered form.

and identify the procedures which contribute most to the slowness. If these procedures implement the technique being compared then our case will be supported. Similarly, a difference in search space might be traced to the two techniques being compared, rather than some more accidental difference between systems. Such micro-analysis is often possible and must often be used to supplement crude statistics.

6 Conclusion

We have argued that experimental work is an indispensable part of ATP work. Theoretical analysis is often more desirable than experimental investigation when both are available, but (a) the theoretical tools are not available to answer the questions we want answered and (b) there are some questions which are inherently experimental and cannot be addressed by theoretical tools.

ATP experimentation is no different, in principle, to experimentation in any science: a hypothesis must be stated and evidence for and against it must be sought and found. Explicit hypothesis statement is not the norm in ATP, but it should become so.

Any experimental work must address two issues: arguing that the results are representative and assigning credit for the results. These issues seem especially difficult in ATP due to the complexity of the problem space and the complexity of the provers. However, there are good methodological practises which can provide more reliable results.

As in all sciences, experimental work should be repeatable by other researchers. By enabling colleagues to carry out similar experiments and analysis, the conclusions can be further confirmed or refuted. In ATP this means making the prover and examples available in some portable form, *e.g.* via ftp, www, disk or tape.

Acknowledgements

Thanks to the JAR editors whose feedback on my earlier proposals inspired some of the ideas above, especially Bob Boyer, Larry Henschen, Rusty Lusk, Michael Richter and Hector Levesque.