Conclusions

about the CADE-13 ATP System Competition

FRANCIS JEFFRY PELLETIER
Department of Computing Science, University of Alberta
Edmonto, Canada
jeffp@cs.ualberta.ca

GEOFF SUTCLIFFE
Department of Computer Science, James Cook University
Townsville, Australia
geoff@cs.jcu.edu.au

CHRISTIAN SUTTNER
Institut für Informatik, TU München
München, Germany
suttner@informatik.tu-muenchen.de

1. Conclusions from the Organizers

The CADE-13 ATP System Competition was clearly a success. The aims for the competition, as set out in the competition design [6], were all met:

- The relative capabilities of the 18 ATP systems that competed were evaluated. The results [5] will help potential users to choose the system that suits their needs.
- The competition stimulated ATP research. In particular, most entrants made special efforts to improve the autonomous performance of their systems, and all the entrants had to ensure that the implementation and debugging of their systems was complete.
- The competition provided an inspiring environment for personal interaction between ATP researchers. (There was more excitment and activity than the organizers expected!)
- Many of the FLoC conference[1] delegates came to see the competition. The competition thus exposed the ATP systems to researchers both within and outside the ATP community.

From an organizational perspective, the competition ran very smoothly, with no technical disasters. This was particularly pleasing for such a "first" event, where little was known about potential problems.

The success of the competition can be attributed to several factors:

[1] CADE-13 was one of four conferences held as part of the Federated Logic Conferences (FLoC) in 1996.

- The careful design of the competition, taking into account community feedback, provided a structure within which all the entrants were happy to participate.
- With one minor exception (see Section 9 of [4] for details), the competition design was faithfully translated into procedures that were used to run the competition.
- Sufficient hardware (supplied by CAIP - the Centre for Aids for Industrial Productivity), useful support software (the SunOS load balancing system and the tptp2X utility), and a community endorsed source of problems (the TPTP problem library), provided an adequate computing environment for the competition.
- The real time presentation of the results during the competition attracted and maintained the interest of many people.
- The efforts and support of many people facilitated various aspects of the competition. Contributors included the ATP community in general, the entrants, the competition organizers[2], the competition panel, the CADE-13 and FLoC organizers, and the CAIP staff.

The competition has had some significant outcomes. For the entrants, their research groups, and their systems, there has been substantial publicity. For the winners, the acknowledgement has been particularly pleasing. The significant efforts that have gone into developing the ATP systems have received public recognition; publications, which adequately present theoretical work, have not been able to expose such practical efforts appropriately. The competition has provided an overview of which researchers and research groups have decent, running, fully automatic CNF ATP systems. All those promises in publications can now be verified against the performances achieved in the controlled environment of the competition.

Of course the competition was not without criticism, both of specific aspects of the CADE-13 competition, and experimental evaluation of ATP systems in general. Some of the major criticisms, and the organizers' comments, are:

- Use of the TPTP for evaluation is biasing research
  Currently no "killer application" is known for ATP. Thus the problems for evaluation cannot come directly from an application domain. The TPTP is large enough in scope and size that good performance on the TPTP probably means good performance in general. The TPTP is continually being expanded, thus preventing the stagnation that would occur using a static problem source.

[2] The first author insists on this acknowledgement of the other two authors, in particular the on site efforts of Geoff Sutcliffe. The organizers in turn would like to acknowledge the efforts of their assistant, Thilo Trapp.

- Time has been wasted on improving implementations
For the future industrial application of ATP systems, reliable, well engineered ATP systems, with short response times, will be required. Developing high quality ATP systems now provides a foundation of techniques that can be used to reach future industrial standards.

- Competitions will damage funding opportunities
This is certainly true if you don't enter the competitions. For winning systems, funding opportunities are more likely to improve than be damaged. Losing systems that show potential, e.g., by solving problems that no other system could, also should receive funding to develop their special capabilities. If a system is completely dominated by others, then indeed funding may be damaged. This does not seem to be a bad thing for the ATP community at large; there is only a limited supply of funding, and useless systems should not be taking money away from the development of better systems. Hopefully funding bodies will have the sense to sustain the funding of projects that are in their infancy, to give them enough time to prove the quality of new ideas in high performance implementations.

- Competitions are not scientific, and give no scientific insight
As is stated in the CADE-13 competition design and results, there was no intention, or claim of ability to, find the best overall ATP systems. The results obtained in a competition must be viewed in the context of the competition design. There are many parameters that affect the results. It does not seem possible to control all the parameters in the way required by formal scientific experiments. Thus competitions cannot claim to be scientific experiments. Rather, competitions should aim to stimulate, motivate, inspire, and expose the entrants and their systems. The success of the annual ACM computer chess championship [3] in this regard shows what can be achieved. In terms of evaluation, competitions do provide some information about the relative capabilities of the systems. They also provide interesting insights into the systems, e.g., which systems appear to be best suited to particular types of problems.

Despite the comments above, the organizers do not claim that the CADE-13 competition was perfect in every detail. The competition was the first such event, and the experience gained will improve future competitions. However, the CADE-13 ATP System Competition was a carefully designed and implemented event, and the benefits by far outweigh the controversial aspects. The evaluation of ATP (and other heuristically guided systems) is a complex task, and there are many alternative points of view (see, e.g., [1] and [2]). As time goes by it is expected that the ATP community will influence and adopt standard evaluation techniques. These techniques can also be used in ATP competitions.

2. Conclusions from the Competition Panel

The competition Panel consisted of Peter Andrews, Alan Bundy, and Jeff Pelletier, and was charged with ensuring that the overall running of the competition was fair. They were also called upon at the last minute to declare one of the competitors to be the recipient of the Kluwer book prize. In the preparation of this document they were also asked to be impartial collectors of comments and suggestions from the competitors about the running of the competition.

2.1. Fairness of Operation of the Competition

The Panel was brought in to provide some independent assurance of the fairness of the competition. We were to make sure that the problems were fairly chosen and presented, and that the competitors all had an equal opportunity to showcase their systems. Finally, the Panel was to rule on any complaints that might arise as the competition proceeded.

In the planning stages of the competition, the main worry concerning possible complaints was that one of the systems might be "cheating" in one way or another. In particular, it was a concern that (a) since only theorems were to be used, and (b) since the competing systems were not required to produce a "proof object" of any sort, a system might simply just "lie" and claim to have proved a theorem when in fact it had not. (This "lying strategy" could be implemented either by having absolutely *no* internal proof system and just claiming to have found a proof after a certain amount of time, or by having something added onto a legitimate prover such that if a proof isn't discovered in a short amount of time then the system claims to have found a proof anyway.) The initial intent was that if any system came under suspicion, the Panel could demand that a "proof object" (or at least enough information from the system so as to be able to construct a proof) be provided off-line for inspection. In actuality, however, the possibility of having to perform this onerous task was eliminated by the simple expedient of requiring contestants to submit the final competition version of their system in advance, and testing this on a number of non-theorems. Now, while it is theoretically possible that a program could have a built-in random generator that randomly claimed to have found proofs – and while it is possible that this generator happened never to have made such a claim in the testing stage but it did produce it in the competition – the Panel thinks this an extremely remote possibility. A less remote possibility (and one which was not looked for) is that a program checks the UNIX system clock and produces an illegitimate claim to have found a proof only if it is the day of the competition. In any case, none of the competitors believed strongly enough that any of the other competitors were "lying" to say anything to the Panel. The Panel does note, however, the possibility of this happening, and recommends that future competitions be aware of it. The Panel also notes that the solution is immediately at hand: A system could and should be required to output a "proof object" for each theorem it claims to have proved. (Alternatively, but perhaps less satisfactorily, the problem set could have some proportion of non-theorems in it.)

The competition specification required that problems of appropriate difficulty be randomly selected from the TPTP library (see [6] for details). The sets of eligible problems were made available to contestants in advance of the competition, and most entrants tested their programs on them. However, due to a programming error, only a subset of the eligible unit equality problems were extracted and transmitted to the contestants. The error was discovered before the competition and the extra problems were included in the final set of eligible problems, but some contestants were not aware that the set of problems had been expanded. Some minor concern has been expressed about this (see Section 2.3 below), but the Panel notes that no contestant got to use this change of problem set to alter their program in advance of the competition, and so it continues to seem fair.

The Panel did not inspect the tptp2X transformations `shorten` and `random`, which changed the presentation of the TPTP problems (see [4] for more details). However, in their final form, the problems used speak for themselves[3]; the predicates, functions, and clause and literal orderings are quite distinct from how they are in the TPTP. The Panel also did not inspect the function which randomly selected the problems to use from the set of eligible problems. The Panel *can* guarantee that it generated a three-digit random seed, and that it appeared to be true that this seed made all the difference in the start-up of the script which randomly selected problems. The Panel has no doubts that this aspect was carried out exactly in accordance with the design specifications.

In closing this subsection on "fairness of the competition", we note that the competition appeared to be fair in the sense that everyone was treated equally. We applaud the organizers for their foresight in designing the various utilities employed: tptp2X, the estimation of problem

[3] The problems are available from the competition WWW site: `http://wwwjessen.informatik.tu-muenchen.de/~tptp/CASC-13/`

difficulty, the start-up script, etc. The actual running of the competitiion went very smoothly due to the forethought of the organizers, especially in their consideration of issues of fairness.

2.2. "Winners", and Decisions made by the Panel

The competition design specified three categories in which there would be "winners". There was a lot of discussion in the year leading up to the competition, and especially in the month or two prior, concerning how a "winner" in any one category would be chosen. Would it be based solely on the number of problems proved? Or should it be based on the total time taken by a system in all its proof attempts (where an unproved problem would garner the maximum allowed time for that problem)? The discussions involved in all this led the organizers to propose two different scoring schemes, both of which are a mixture of the two considerations.

It turns out that both scoring schemes yielded the same winners, for each category. They also identically ranked all the systems, within each category. Nonetheless, there was much agreement among the entrants that the number of problems proved should be the sole determinant of winning, using total times only to break ties. The Panel notes that this has been incorporated in the CADE-14 competition. Given that the two scoring schemes for winning in a category and the fact that the two scoring schemes yielded the same winners, there seems to be little reason to think that any system was unfairly treated with regard to whether it was or wasn't a winner in one of the three categories.

There are two other topics that raised considerable discussion. The first of these topics was about the choice of categories for the competition. Here there are three concerns that were raised. First, some competitors felt that there should have been a category of problems without equality, since some provers are not designed to handle equality but are very good on non-equality problems. The Panel notes that this will happen in the CADE-14 competition. Secondly, many people (both competitors and others) thought there should have been non-theorems added into the mix of problems. This would have given model-generation and clause-linking theorem provers at least some chance to display their strengths. Although its omission from the competition is not unfair since this was a well-advertised feature of the competition, the Panel is pleased to note that the CADE-14 competition will include such a category. The third concern within the topic was the restriction to negated conclusion clausal form. This format for presenting problems essentially restricts the class of competing systems to resolution provers. The Panel is pleased to note that a category of natural form, full first order problems, is in the CADE-14 competition.

The second topic that generated considerable discussion concerned the last minute donation of a book prize by Kluwer Press. Since it was last minute, there was no prior discussion about how to determine the recipient of this prize. It therefore fell to the Panel to decide, using whatever features of the competition and its results that they individually felt to be relevant, which system should receive this prize. We should make it clear that we did not view this recipient as some sort of "overall winner", because each of the categories was different. Instead we tried to look at additional qualities – e.g., how many problems were solved *really* fast, or how robust it seemed to be, or even what language it was written in – besides comparing it with other provers from other categories (to the limited extent that this is possible).

In the end, E-SETHEO's performance in the mixed-monolithic class carried the day, in the (qualitative) judgment of the Panel. The reasons for this include the following: In the initial design of the categories for the competition, it was thought that "compositional" systems would have a major advantage over "monolithic" systems (see [6]). Of course, a monolithic system *could* have entered the compositional category, and in fact the problems given to these two the categories were identical. Despite the fact that E-SETHEO did not switch strategies, it nonetheless outperformed the winning system in the compositional category. And despite the fact that it was not entered in the unit equality category (and the Panel thinks it would have done rather poorly in that), the Panel was quite impressed that a monolithic system did so well in comparison to the compositional systems. And so the Panel decided to give the Kluwer book prize to E-SETHEO.

Other systems also caught the Panel's eye, but just not so strongly. It was thought even that a system that did not win any category might nonetheless be "strong" or "noteworthy" enough to win. But in the end, the Panel decided that E-SETHEO was even more noteworthy. But we should remark that a version of Otter was entered in all three categories: a version in "autonomous mode" entered the compositional and unit equality categories, and a version with preset switches chosen by Larry Wos entered the monolithic category. Otter won the unit equality category, and the versions of Otter entered in the other two categories each took second place. It certainly is an impressive showing that one system can be configured so as to run, and run well, in many different categories, and the Panel was indeed impressed. But in the end they decided in favor of E-SETHEO nonetheless, for the reasons mentioned above.

The Panel worries that there might be a general perception that winning the Kluwer book prize is interpreted as being “an overall winner”. The description above should show that such an interpretation is not at all justified or intended. There was, in the CADE-13 competition, no such overarching category, but instead there were three subcategories in which there were winners.

2.3. Comments from Contestants (and others) on the CADE-13 (and future) Competitions

The competition organizers asked the Panel to act as a neutral collection agency for comments and criticisms of the competition from both competitors and other interested observers. The most common and most emphasized comments were gratitude that this competition took place and appreciation of the efforts of the organizers for dealing effectively and seriously with the many items large and small that came up in the planning and execution of the competition. The Panel joins with these competitors in expressing their sincere thanks from the automated reasoning community.

We have already above seen some of the comments made by competitors (and others) concerning such topics as the desirability of producing a “proof object”, of including non-theorems into the competition, of making the number of problems correctly solved be the measure of success (and using times only to break ties), of having a category of problems without equality, and of having a category of natural form, first order logic.

We also remarked on the considerations that went into selecting the recipient of the Kluwer book prize. The fact that this was a last-minute addition meant that the criteria relevant to winning this prize could not be publicized in advance. And this in turn led to many misapprehensions, and a lot of email, as to what the prize meant and how the recipient was chosen. The Panel feels that there is nothing inherently wrong with choosing a prize winner on the basis of “qualitative” and perhaps subjective criteria. But it would be a service if the fact that such considerations were going to be used was made known in advance. The Panel thinks that there could even be an “overall winner” chosen on these types of grounds – but then this ought to be made known in advance and the Panel should be prepared to discuss their choice and reasons, both among themselves in advance and publically after the decision is made.

Other comments that were received by the Panel can be seen as falling into a variety of different categories.

- Comments about "attitude":
  "The contest should be taken in a spirit of fun and the results should not be taken too seriously. It gives a limited view of the performance of the provers."

  "The competition was a very enjoyable 'sportive event'. For the next competition there might be a split between a 'sportive event' like the CADE-13 competition with similar conditions and a serious 'statistical event'."
- Comments about the problem set:
  "We need a wider, more diverse problem set in order to make the competition really significant for the evaluation of practical theorem provers."

  "It is possible that the TPTP problem set is not representative of the theorems one might find in various applications or even in general. Perhaps it is not a valid set of problems to use in testing."[4]
- Comments about categories in the competition (besides those above):
  "The monolithic and compositional categories should be combined, especially since there seems to be little advantage to compositional systems based merely on the fact that they're compositional."
- Comments about evaluation:
  "The data collected during the competition is dominated by 30% timeouts and 25% "one second" runs. So the statistical evaluation that has been done is somewhat speculative."[5]

  "There ought to be also a 'serious statistical event' taking more time (e.g., several weeks in total) to evaluate more problems (e.g., all the eligible problems) with much higher timeouts (e.g., 3600 seconds)."
- Changing problems:
  "The only thing we did not like was that a lot of problems were added to the set of eligible problems a couple of days before the competition started. We had used the original set to find a general purpose heuristic, and it failed on the new problems ... not because of a weakness of the program but solely because of the chosen heuristic."[6]
- Practical issues:
  "There were some practical problems that were totally out of the organizers' control, but might be worth mentioning as issues to consider for future contests. I had trouble accessing the contest machine (to

[4] This issue is discussed in Section 1 above.

[5] This comment is specific to the unit equality category.

[6] This comment came from an entrant in the unit equality category, and this issue is discussed in Section 2.1 above.

install my system) over the net from where I live. It was usually an incredibly slow connection. Also, there was a problem with software availability on the contest machine ..."

- Other:
  "Those who run the competition or design it should not have an entry in the competition. This is a conflict of interest and can introduce all kinds of bias, from the choice of the problems, to the scoring function, to having an advantage of advance knowledge. This can happen in many subtle ways, no matter how sincere one tries to be."

3. Future ATP System Competitions

The CADE-13 ATP System Competition was the first large scale controlled competition for 1st order ATP systems. Future competitions should extend the competition scope. The two most important issues are to extend the competition to FOF (First Order Form) and model generation problems. Within the details of the competition organization, finer grained competition categories determined by problem characteristics will make it possible to find specialist systems for specific types of problems. This also allows a system that is specialized for a particular problem class to win the corresponding category, even if its overall performance is poor. Conversely, generalists can be also be recognized even if they do not win any category. Entrants must be encouraged to have their systems evaluated in as many competition categories as is meaningful, in order to determine both the strengths and the weaknesses of the systems. This increases the value of competitions as scientific experiments.

In the face of the desirability of more extensive competitions, it is also important to avoid being overambitious. A successful limited competition with meaningful results is preferred over a larger competition which is not accepted by the ATP community. Large events require large resources, which are currently not available. If the CADE ATP competitions continue to provide useful information about the competing systems, it would be reasonable for those who benefit from access to the information to help provide the resources required for bigger and better competitions.

In conclusion ... after more than 40 years of research, automated theorem proving abounds with techniques developed and ideas proposed. The future requires that these techniques and ideas be evaluated to determine which are viable, and to integrate them into systems which are far more flexible and powerful than the systems available today.

The most important challenge ahead in theorem proving research will be to provide adequate control - a subject still in its infancy, since it is difficult to approach theoretically.

For all these goals, experimental system evaluation is a crucial research tool, and competitions will provide stimulus and insight that can lay the basis for the development of future ATP systems.

References

1. J.N. Hooker. Testing heuristics: We have it all wrong. *Journal of Heuristics*, 1:33–42, 1996.
2. D.S. Johnson. A Theoretician's Guide to the Experimental Analysis of Algorithms. Draft available by ftp from dimacs.rutgers.edu:pub/dsj/temp/exper.ps, 1996.
3. D. Kopec, M. Newborn, and M. Valvo. The 22nd Annual ACM International Computer Chess Championship. *Communications of the ACM*, 35(11):100–110, 1992.
4. G. Sutcliffe and C.B. Suttner. The Procedures of the CADE-13 ATP System Competition. *Journal of Automated Reasoning*, This issue, 1997.
5. G. Sutcliffe and C.B. Suttner. The Results of the CADE-13 ATP System Competition. *Journal of Automated Reasoning*, This issue, 1997.
6. C.B. Suttner and G. Sutcliffe. The Design of the CADE-13 ATP System Competition. *Journal of Automated Reasoning*, This issue, 1997.