Request for Feedback/Comments

Please create a pull request or an issue (strongly preferred) or email Daniel Le Berre directly at leberre@cril.fr

---

The SAT Practitioner's Manifesto v1.0

Practical SAT research is both time and resource consuming. As such, it is important that the community provides some guidelines to make sure that all efforts made by practitioners are properly recognized during paper evaluation.

The following principles are adopted by the community.

1. Benchmarks should be available for research purposes.
2. Solvers should be available in binary form for research purposes.
3. A recent generic benchmark set (e.g. competition benchmarks) should be chosen among those of the last 3 years.
4. Experimental results should include a comparison with the state of the art.
5. Details on the experimental conditions should be provided (e.g. hardware, OS).

The authors may indicate the number of principles followed in their evaluation so as to signal to the reviewers of the soundness of the methodology. For example with phrases such as "Our evaluation follows 5/5 principles laid out in SAT mainfesto (version x.xx)".

Benchmarks should be publicly available for research purposes

Papers presenting results on benchmarks should provide a link to those benchmarks, preferably the raw benchmarks, else a generator for those benchmarks.

In general, the community is also interested in results that improve SAT solving on non-public benchmarks. Still authors are expected to make their research reproducible and extensible by others, by providing for instance artificial but public benchmarks, with similar characteristics, on which the new methods are also evaluated.

Solvers should be available in binary form for research purposes

Papers presenting experimental results about a system (solver) should allow the community to access this solver in at least binary form, or even better in source form. This allows to check performance claims independently.

A recent generic benchmark set (e.g. competition benchmarks) should be chosen among those of the last 3 years

It takes time and computer resources to collect empirical evidence and to dig details about benchmarks sets. All data retrieved on a set of benchmarks from a competitive event in year X should not be void as soon as a new set of benchmarks is published. As such, the community acknowledges that the entire benchmark set from year X will be acceptable until year X+3 included.

Experimental results should include a comparison with the state of the art

The community has established regular competitive events to publicly promote state-of-the-art systems. It is expected that experimental results, about benchmarks or solvers, include state-of-the-art systems highlighted during recent competitive events. The comparison could also include reference systems (widely used systems which might no longer be considered state of the art).

It is also useful to choose the experimental setup in such a way that a comparison with other published data is easy. One could use similar hardware, corresponding time limits, if necessary provide scaling factors, or at least run reference solvers too. For instance if the authors chose to run the benchmark from a certain competition, then the evaluation will be more instructive if results for another solver winning or at least participating in that competition are also provided, unless of course, the time-limit and the hardware setup are directly comparable.

In order to better understand the contribution to the state of the art, and since competition results provide a large resource of independent experiments, if the benchmark suite from year X is used, then it is advisable to include a comparison to top performing solvers for all the years since (including) X.

> Many papers report results on arbitrarily merge benchmarks and then only report results on those. Then one can not check out other papers (or competition results), to figure out what this really means (without running the competition again on the benchmarks). It is better to use a canonical set of benchmarks, for which the community already has numbers. So using only 2018, 2019, 2020 benchmarks alone make sense, but not merging them (or some pairs) and ONLY reporting on the merged instance set. The point is that a reader can compare the results with existing results without the need to run something (or do some complicated spreadsheet hack).

Besides reporting on the number of solved instances or run-time (including PAR2 scores etc.) it is most instructive to also include secondary statistics. For instance if a new method is supposed to reduce the number of conflicts, then of course the number of conflicts should be included.

Details on the experimental conditions should be provided (e.g. hardware, OS)

The experimental results much depend on the hardware (CPU, amount of RAM, L2/L3 cache, NUMA architecture, etc.) and operating system. Those details should be provided when reporting the results.

Signatories

- Daniel Le Berre (Université d'Artois/CNRS)