Lemma Mining over HOL Light

Cezary Kaliszyk and Josef Urban

1 University of Innsbruck, Austria
2 Radboud University, Nijmegen

Abstract. Large formal mathematical libraries consist of millions of atomic inference steps that give rise to a corresponding number of proved statements (lemmas). Analogously to the informal mathematical practice, only a tiny fraction of such statements is named and re-used in later proofs by formal mathematicians. In this work, we suggest and implement criteria defining the estimated usefulness of the HOL Light lemmas for proving further theorems. We use these criteria to mine the large inference graph of all lemmas in the core HOL Light library, adding thousands of the best lemmas to the pool of named statements that can be re-used in later proofs. The usefulness of the new lemmas is then evaluated by comparing the performance of automated proving of the core HOL Light theorems with and without such added lemmas.

1 Introduction

In the last decade, large formal mathematical corpora such as the Mizar Mathematical Library [5], Isabelle/HOL [33] and HOL Light [7]/Flyspeck [6] have been translated to formats that allow easy experiments with external automated theorem provers (ATPs) and AI systems [10, 17, 26]. Several AI/ATP methods for reasoning in the context of a large number of related theorems and proofs have been suggested and tried already, including: (i) methods (often external to the core ATP algorithms) that select relevant premises (facts) from the thousands of theorems available in such corpora [8, 15], (ii) methods for internal guidance of ATP systems when reasoning in the large-theory setting [31], (iii) methods that automatically evolve more and more efficient ATP strategies for the clusters of related problems from such corpora [28], and (iv) methods that learn which of such specialized strategies to use for a new problem [14].

In this work, we start to complement the first set of methods – ATP-external premise selection – with lemma mining from the large corpora. The main idea of this approach is to enrich the pool of human-defined main (top-level) theorems in the large libraries with the most useful/interesting lemmas extracted from the proofs in these libraries. Such lemmas are then eligible together with (or instead of) the main library theorems as the premises that are given to the ATPs to attack new conjectures formulated over the large libraries.

This high-level idea is straightforward, but there are a number of possible approaches involving a number of issues to be solved, starting with a reasonable definition of a useful/interesting lemma, and with making such definitions efficient over corpora that contain millions to billions of candidate lemmas. These
issues are discussed in Sections 4 and 5 after motivating and explaining the overall approach for using lemmas in large theories in Section 2 and giving an overview of the recent related work in Section 3.

As in any AI discipline dealing with large amount of data, research in the large-theory field is driven by rigorous experimental evaluations of the proposed methods over the existing corpora. For the first experiments with lemma mining we use the HOL Light system, together with its core library and the Flyspeck library. The various evaluation scenarios are defined and discussed in Section 6 and the implemented methods are evaluated in Section 7. Section 8 discusses the various future directions and concludes.

2 Using Lemmas for Theorem Proving in Large Theories

The main task in the Automated Reasoning in Large Theories (ARLT) domain is to prove new conjectures with the knowledge of a large body of previously proved theorems and their proofs. This setting reasonably corresponds to how large ITP libraries are constructed, and hopefully also emulates how human mathematicians work more faithfully than the classical scenario of a single hard problem consisting of isolated axioms and a conjecture [30]. The pool of previously proved theorems ranges from thousands in large-theory ATP benchmarks such as MPTP2078 [1], to tens of thousands when working with the whole ITP libraries.

The strongest existing ARLT systems combine variously parametrized premise-selection techniques (often based on machine learning from previous proofs) with ATP systems and their strategies that are called with varied numbers of the most promising premises. These techniques can go quite far already: when using 14-fold parallelization and 30s wall-clock time, the HOL(y)Hammer system [10][11] can today prove 47% of the 14185 Flyspeck theorems [12]. This is measured in a scenario in which the Flyspeck theorems are ordered chronologically using the loading sequence of the Flyspeck library, and presented in this order to HOL(y)Hammer as conjectures. After each theorem is attempted, its human-designed HOL Light proof is fed to the HOL(y)Hammer’s learning components, together with the (possibly several) ATP proofs found by HOL(y)Hammer itself. This means that for each Flyspeck theorem, all human-written HOL Light proofs of all previous theorems are assumed to be known, together with all their ATP proofs found already by HOL(y)Hammer, but nothing is known about the current conjecture and the following parts of the library (they do not exist yet).

So far, systems like HOL(y)Hammer (similar systems include Sledgehammer/MaSh [13] and MaLARea [29]) have only used the set of named library theorems for proving new conjectures and thus also for the premise-selection learning. This is usually a reasonable set of theorems to start with, because the human mathematicians have years of experience with structuring the formal libraries. On the

1 14185 theorems are in the HOL/Flyspeck library, about 20000 are in the Isabelle/HOL library, and about 50000 theorems are in the Mizar library.

2 A similar scenario has been introduced in 2013 also for the CASC LTB competition.