Theory Reduction, Theory Revision, and Retranslation

Allen Ginsberg
AT&T Bell Laboratories
Holmdel, NJ 07733
abg@vax135.att.com

Abstract
This paper presents an approach to retranslation, the third and final step of the theory reduction approach to solving theory revision problems [3,4]. Retranslation involves putting a modified “operationalized,” or “reduced,” version of the desired revised theory back into the entire language of the original theory. This step is desirable for a number of reasons, not least of which is the need to “compress” what are generally very large reduced theories into much smaller, and thus, more efficiently evaluated, unreduced theories. Empirical results for the retranslation method are presented.

Introduction and Overview
A theory revision problem exists for a theory \( T \) when \( T \) is known to yield incorrect results for given cases in its intended domain of application. The goal of theory revision is to find a revision \( T' \) of \( T \) which handles the set of all known cases correctly, makes use of the theoretical terms used in \( T \), and may, with a reasonable degree of confidence, be expected to handle future cases correctly.

This paper is about retranslation: the third, and final, step of the theory reduction approach to solving theory revision problems. The first step of the approach, discussed in detail in [3], is to “translate” the theory in question into a form that is more amenable to inductive learning techniques. This may be viewed as a complete prior “operationalization” of the theory, in the sense of the term employed in explanation-based learning [7]. The resulting translation is called the reduced theory because the number of distinct primitive terms employed by this theory is fewer than that of the original. In terms of the number of statements (distinct clauses or rules) it contains, however, the reduced theory will generally be much larger than its unreduced counterpart. The second step of the approach, presented in [4], involves modifying the reduced theory in order to improve its ability to “give the correct answer” relative to the given set of cases, \( C \), but in such a way that it is reasonable to expect improved performance over cases not included in \( C \) as well. RTLS (Reduced Theory Learning System) is the system that performs this step. Once the reduced theory has been modified to cover all the cases in \( C \), the final step involves a “retranslation” of the modified reduced version back into the entire language of the original theory. This step is necessary/desirable for a number of reasons, one of them being the desire to “compress” what are generally very large reduced theories into much smaller, and thus, more efficiently evaluated, unreduced theories.

In the previously cited papers I asserted that 1) reduction of non-trivial medium-sized expert systems theories could be achieved in acceptable times, 2) good improvements in performance could be achieved by training the reduced theory using the methods discussed in [4], and 3) that a method for automatic retranslation of expert system theories was known. While the first two assertions were, and still are, justifiable, assertion (3), as I stated in [2], was premature: it turned out that the simple retranslation algorithm I had in mind would actually produce an egregiously overgeneralized result. My initial suspicion that retranslation would be a difficult problem, even for expert system theories, was in fact correct.

Thus the raison d’être of this paper: to present recent research results on the retranslation problem, and in so doing to present a sound approach for doing retranslation. First, however, after describing the problem in detail in the next section, it will be shown that the notion of retranslation, properly understood, is a problem for theory revision in general, as well as other AI endeavors.

Problem Statement

Theory Reduction
Theories posit inferential connections leading from “observable features” characteristic of some class of phenomena, to collections of theoretical terms that have explanatory and/or predictive power with respect to systems that exhibit these features. Theory reduction is essentially a matter of compilation of the evidential relations holding between observables and theoretical terms in a theory, and is not intended to carry the ontological or semantical connotations associated
Table 1: Some Symbols and Terminology Defined

- **Answer**($c$): the given (correct) theoretical description for case $c$; may contain several t-terms.
- **$\tau$-case**: a $c$ whose Answer($c$) includes $\tau$. **Non-$\tau$-case**: a $c$ whose Answer($c$) does not include $\tau$.
- **Rules-for($\tau$)**: the set of rules in theory $T$ that directly conclude $\tau$. **Level($\tau$)**: max of the levels of Rules-for($\tau$)
- **Endpoint**: a t-term that does not occur in the antecedent of any rule.
- **RTLS-label($\tau$)**: the label generated for endpoint $\tau$ by the RTLS system.
- **Label(e)**: the set of minimal environments generated by calculating the label of expression $e$ (a conjunction of t-terms and observables), where every t-term in $e$ has its original label or is assigned some $\delta$-label.
- **Rule-correlated-t-terms($\tau$)**: the set of all t-terms occurring in some member of Rules-for($\tau$).
- **Rule-correlated-observables($\tau$)**: the set of all observables occurring in some member of Rules-for($\tau$).
- **Theory-correlated-t-terms($\tau$)**: the set of all t-terms that occur in any rule that is a “link” in a “rule-chain” having some rule in Rules-for($\tau$) as the last link.
- **Theory-correlated-observables($\tau$)**: the set of all observables that occur in any rule that is a “link” in a “rule-chain” having some rule in Rules-for($\tau$) as the last link.

with reductionism in natural science or certain philosophical movements [5].

Let $T$ be the theory and let the vocabulary (predicate symbols, propositional constants, etc.) of $T$ be divided into two disjoint subsets $T_0$ and $T_1$. We refer to these as the observational (operational) and theoretical (non-operational) vocabulary of $T$, respectively [8]. Let $\tau$ be a member of $T_1$, and let $o_1 \ldots o_k$ be a conjunction of distinct items where $o_i \in T_0$ for $i = 1, \ldots, k$. Suppose that the statement $o_1 \ldots o_k \rightarrow \tau$ follows from $T$. Moreover, suppose that if any conjunct is removed from $o_1 \ldots o_k$ this would not be the case. Then $o_1 \ldots o_k$ is a minimal sufficient purely observational condition for $\tau$ relative to $T$. Now let $O_\tau$ represent the set of all conjunctions $o_1 \ldots o_k$ such that $o_1 \ldots o_k$ is a minimal sufficient purely observational condition for $\tau$ relative to $T$. Then $O_\tau$ is the called the reduction of $\tau$ with respect to $T$. Following the terminology of de Kleer [1], we sometimes call $O_\tau$ the label for $\tau$, and each member of $O_\tau$ is said to be an environment for $\tau$. The set of all $O_\tau$ for $\tau \in T_1$, denoted by $R(T)$, is called the reduction of the theory $T$.

Reduction of Expert System Theories

We consider an expert system theory $E$ to be a restricted propositional logic theory. That is, $E$ consists of a set of conditionals in propositional logic, i.e., the rules or knowledge base. A sentence $\alpha \rightarrow \beta$ is considered to follow from $E$ iff, to put it loosely, $\beta$ can be derived from $\alpha$ and $E$ via a sequence of applications of a generalized version of modus ponens. $E$ is said to be acyclic if, roughly speaking, a sentence of the form $\alpha \rightarrow \alpha$ does not follow from $E$.

In [3] I presented a two-step algorithm for the complete prior reduction of acyclic expert system theories, and discussed a system, KB-Reducer, that implements the algorithm. In the first step the rules in $E$ are partitioned into disjoint sets called rule levels. A rule $\tau$ is in level 0 iff the truth-value of the left-hand side of $\tau$ is a function of the truth-values of observables only. A rule $\tau$ is in level $n$, iff the truth-value of the left-hand side of $\tau$ is a function of the truth-values of observables and theoretical terms that are concluded only by rules at levels 0, . . . , $n - 1$. This partition defines a partial-ordering for computing the reduction of all theoretical terms: each rule in level 0 is processed (exactly once), then each rule in level 1, etc. For further details see [3].

Retranslation as a General Problem

The subject of this paper is called ‘retranslation’ in relation to the aforementioned reduction process which may be termed a ‘translation’ of a theory into a form which avoids the use of theoretical terms (on the left-hand-sides of rules). In retranslation we are interested in re-expressing knowledge currently expressed solely in “low-level” observational terms, in more compact “high-level” theoretical terms. The problem of reinterpreting/reassimilating low-level data or results in terms of high-level constructs is a key aspect of many AI problems, e.g., vision.

As a concrete example in the domain of theory revision consider the case of Kepler’s laws of planetary motion in relation to Newton’s laws of motion (including the law of gravitation). Kepler’s laws are an example of what philosophers of science call empirical generalizations [8], i.e., statements couched solely in terms of observables, e.g., planet, elliptical orbit, sun. Newton showed that these laws are consequences of his laws of motion, which involve the theoretical notions of force and gravitational force. That is, from Newton’s laws, together with certain necessary “auxiliary” statements, e.g., a planet is a massive body, Kepler’s laws can be derived. Thus, if one were to reduce the Newtonian theory, one would find Kepler’s laws (or a set of more primitive statements equivalent to them)
in the reduced theory. Note that this reduced theory
would not contain theoretical terms such as force and
gravity. The process of restating this reduced theory
- which would contain Kepler's laws and other purely
observational statements - in terms of a theory that
posits unobservables, is retranslation.

Relaxed Retranslation
While we may consider Kepler's laws to be part of the
reduction of Newton's theory, it is not correct to sug-
gest that Newton had, in any sense, the entire reduc-
tion of his theory (or a variant thereof) available to
him prior to its formulation in theoretical terms. New-
ton's laws entailed empirical generalizations that were
not predicted and verified until well after the formu-
lation of the theory. This illustrates the idea that the
notion of retranslation - re-expressing something at a
theoretically richer conceptual level - and the notion of
generalization - formulating a more powerful version of
something already known - cannot, in practice, be en-
tirely divorced from one another. Thus one answer to
the question, Why retranslate?, is that this is simply
another way of trying to broaden our knowledge. To
help clarify the meaning and pertinence of this point
of view consider the following points.

In general, it is likely that a retranslation problem
will start with a reduction $R(T')$ for a theory $T'$ that is
not the same as the reduction of the "ultimate desired
version" of the theory. For most intents and purposes,
it is reasonable to assume that this will be the case
for all but very small "toy" theories. For this reason
it seems foolish to insist that the retranslation process
should necessarily yield a theory whose own reduction
is exactly identical to the given reduction $R(T')$. In-
stead of viewing $R(T')$ as an absolute constraint on the
result - to be preserved at all costs - we should view
it as providing guidelines on the retranslation process.
We call this version of the problem relaxed retransla-
tion. It is this version of retranslation that is most
similar to the Newton-Kepler example. All that is re-
quired of the generated retranslation is that its per-
fomance over the cases $C$ be at least as good as that of
$R(T')$. In the sequel it is this version of the retransla-
tion problem that we will address.

Finally, we should note a way in which the Newton-
Kepler example differs from the retranslation problems
addressed here. While Newton undoubtedly had some
notion of force as part of the received knowledge of the
time, the fact is that he really can be said to have in-
vented this theoretical concept, and others, because
he formulated precise laws that governed their use.
In contrast, the retranslation problems addressed by this
work always take place within the context of a set of
given theoretical terms, and an initial, albeit flawed,
version of the theory. As we will soon see, the struc-
tural relationships among the various components of
the theory - as embodied in its rules - provides cru-
cial information in helping to guide the search for suit-
able retruns. Recognizing the need {utility) for
{of} new theoretical terms, while a relevant avenue for
future investigation, is a task that is not directly ad-
dressed by the methods presented here.

Retranslation
As with any technical topic, one needs to introduce
a certain amount of terminology in order to keep the
presentation brief and precise. To make for easier ref-
ence, most of the special vocabulary used in this pa-
per is defined in Table 1. A number of these ideas, in
particular, the crucial notions of rule-correlated and
theory-correlated observables and theoretical-terms,
are illustrated in the example in Figure 1.

Top-Down Retranslation
Let $R$ be the reduction we wish to retranslate, and
let $T$ be the version of the theory we were given prior
to the learning session. For every endpoint $\tau \in \mathcal{T}_\tau$ -
where $\tau$ is an endpoint if it does not occur in the
left-hand-side of any rule in $T$ - a corresponding RTLS
label, RTLS-label($\tau$), will exist in $R$ (this is the output
of RTLS). Since the original theory was acyclic some
endpoints must exist. In top-down retranslation we
start with endpoints: for a given endpoint, $\tau$, we first
try to find changes in the Rules-for($\tau$) and in the la-
bes of the theoretical terms, t-terms for short, in these
rules so that the label for $\tau$ generated by these changed
rules and labels is either identical to, or fairly close to,
RTLS-label($\tau$). What is important is that this label
generated for $\tau$ - we call it a $\delta$-label - yields the same
performance results over the cases as RTLS-label($\tau$).

Intuitively this process corresponds to asking the
question: What would the labels of the t-terms used
to conclude $\tau$ - given by Rule-correlated-t-terms($\tau$) -
as well as the new Rules-for($\tau$) have to "look like," in
order for RTLS-label($\tau$) (or something "close enough"
to it) to be the label that the retranslated theory will
 generate for $\tau$? Suppose that we have answered this
question to our satisfaction: then we have succeeded in
pushing, or, to borrow a phrase, "back-propagating,"
the retranslation problem for $\tau$, down one level of the
theory. Let $\lambda$ be any member of Rule-correlated-t-
terms($\tau$). Now the question is: What would the labels of
the t-terms in Rule-correlated-t-terms($\lambda$) and the
new Rules-for($\lambda$) have to look like in order for the $\delta$
label of $\lambda$ to be generated?, and clearly we have to ask
this question for every $\lambda \in$ Rule-correlated-t-terms($\tau$).
We continue to ask this question all the way down the
rule levels until we reach the zeroth level. Since rules
at the zeroth level make use solely of observables on
their left-hand-sides, we will know exactly what the
rules at this level should look like: if $\tau$ is a t-term at
this level the new Rules-for($\tau$) will come directly from
the $\delta$-label($\tau$) generated by the top-down retranslation
procedure.

While the general idea sounds simple enough, there
are, in fact, many ways in which things can fail to
go smoothly. In order to focus ideas we will look at a small, but representative, example in some detail; Figures 1 and 2 are used to illustrate this example.

We proceed on an endpoint by endpoint basis, i.e., we solve the retranslation problem for one endpoint and then move on to another. Every endpoint requiring retranslation, i.e., every endpoint that has an RTLS-label different from its original label in $T$, will be processed once and only once. This immediately raises the question of "interactions" among endpoints that share theory-correlated $t$-terms. For example, in requiring retranslation, i.e., every endpoint that has an

\[ \text{Rule-correlated-observables}(T) \]

that effect more than one endpoint. $T$-terms that are theoretically-correlated to a single endpoint are called eigen-terms. We proceed on an endpoint by endpoint basis, i.e.,

\[ \text{Rule-label}(r) \]

where these are the observables and $t$-terms that occur in some rule that directly concludes $T$. This means that we have a $\delta$-label($\tau$) at this point (either RTLS-label($\tau$) if $\tau$ is an endpoint, or else the current $\delta$-label for $\tau$ as determined by the retranslation of the endpoint(s) to which $\tau$ is theoretically-correlated). We begin by identifying Rule-correlated-observables($\tau$) and Rule-correlated-$t$-terms($\tau$), where these are the observables and $t$-terms that occur in some rule that directly concludes $\tau$. We now try to "interpret" or "reconstruct" $\delta$-label($\tau$) by finding a set of rules for $\tau$ using these items as components. That is, for each environment $e = o_1 \ldots o_n \in \delta$-label($\tau$), we attempt to partition the observables in $e$ into sets corresponding to the various "contributions" that would be made by some rule containing these components. These rules are said to be interpretations of the environments that generate them. For example, in Figure 2, we see that each environment of the RTLS-label for $\tau_3$ can be viewed as arising from the rule $n\tau_3 \rightarrow \tau_3$ provided that the appropriate modifications to the label of $\tau_3$ are made. In this Figure parentheses and bold-face are used to indicate the portion of the interpreted environment that is being "accounted for" by the indicated $t$-term. For example, in the interpretation $n\tau_3$ (abel), abel is the portion of abeln coming from $\tau_3$.

There are three activities included in the interpretation-forming phase. In the first place we are generating candidates for the new Rules-for($\tau$). The "external structure" of these rules can be identical to rules in the original theory, or they may generalize and/or specialize these rules in certain ways. In the second place we are determining the content of the $\delta$-labels of the $t$-terms that are

<table>
<thead>
<tr>
<th>$t$-term</th>
<th>Level</th>
<th>Label in $T$</th>
<th>Correlated Observables</th>
<th>Correlated Theory Observables</th>
<th>Correlated Theory $t$-terms</th>
<th>Correlated Theory Eigen-Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_1$</td>
<td>0</td>
<td>$ab \lor ac \lor bc$</td>
<td>$a, b, c$</td>
<td>$a, b, c$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\tau_2$</td>
<td>0</td>
<td>$ad \lor ae \lor de$</td>
<td>$a, d, e$</td>
<td>$a, d, e$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\tau_3$</td>
<td>1</td>
<td>$abfg \lor acfg \lor bcfg \lor adf \lor adl \lor ael \lor edl$</td>
<td>$f, g, l$</td>
<td>$\tau_1, \tau_2, \tau_3$</td>
<td>$a, b, c, d, e, f, g, l$</td>
<td>$\tau_1, \tau_2, \tau_3$</td>
</tr>
<tr>
<td>$\tau_4$</td>
<td>1</td>
<td>$adh \lor aeh \lor deh \lor dhk$</td>
<td>$h, k$</td>
<td>$\tau_2, \tau_3, \tau_6$</td>
<td>$a, d, e, h, k$</td>
<td>$\tau_2, \tau_6, \tau_6$</td>
</tr>
<tr>
<td>$\tau_5$</td>
<td>2</td>
<td>$abfgn \lor acfgn \lor bcfn \lor adln \lor aeln \lor deln$</td>
<td>$n$</td>
<td>$\tau_5$</td>
<td>$a, b, c, d, e, f, g, h, l$</td>
<td>$\tau_1, \tau_2, \tau_3, \tau_1, \tau_3$</td>
</tr>
<tr>
<td>$\tau_6$</td>
<td>0</td>
<td>$d \lor h$</td>
<td>-</td>
<td>$d, h$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\tau_7$</td>
<td>0</td>
<td>$ce$</td>
<td>-</td>
<td>$c, e$</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\tau_8$</td>
<td>1</td>
<td>$cex \lor cey$</td>
<td>$x, y$</td>
<td>$\tau_7$</td>
<td>$c, e, x, y$</td>
<td>$\tau_7, \tau_7$</td>
</tr>
</tbody>
</table>

**Figure 1**

Endpoints of $T$: $\tau_4, \tau_5, \tau_6$

RTLS-label($\tau_4$): $adef \lor dhk$

RTLS-label($\tau_5$): $abeln \lor abeln \lor bdeln \lor abfn \lor begn$

RTLS-label($\tau_6$): $cex \lor cey$
used to conclude $\tau$. Consider, for example, the re-
translation of $\tau_3$ in Figure 2. In this case each desired
environment happens to generate the same interpre-
tation $n_{\tau_3}$ (which is, in fact, identical to a rule in the
original theory), but each environment "impacts" a dif-
ferent environment from the original label of $\tau_3$. For
example, the desired environment $abeln$ forces a spe-
cialization of the environment $ael$ in the original label
to the environment $abc$ in the new label, while the
environment $bcgn$ forces a generalization of the envi-
ronment $bcfg$ in the original label to $bcg$ in the new
label. Therefore, in the third place, we have to make
sure that the new label that is generated for $t$-terms,
$\tau_3$ in the example, accurately reflects all the changes
arising from the interpretations that are, at least tenta-
atively, being considered. In the current system this
is achieved by obeying the following regimen. We first
perform all the specialization modifications to the ori-
ginal label. Whenever we add a specialized environment
e we must be sure to remove all the environments that
are more general than e from the label. We then per-
form all the generalization modifications. Finally, we
re-minimize the resulting label.

There are two main complications that can occur in
the interpretation-forming phase. It simply may be im-
possible to interpret all the environments of $\delta$-label($\tau$)
in terms of the items in Rule-correlated-observables($\tau$)
and Rule-correlated-$t$-terms($\tau$). This will certainly
be the case if some $e \in \delta$-label($\tau$) contains one or
more observables that are not in theory-correlated-
observables($\tau$).

In fact it is easy to know in advance whether or not
theory-correlated-observables($\tau$) will have to be aug-
mented with new observables in the new theory. A
simple criterion is the following: if there are two cases
c_1, c_2, one a $t$-case, and the other not, such that c_1, c_2
share exactly the same theoretically-correlated observ-
able of $\tau$, then we know that we will have to make use
of the other observables in these cases if we are to con-
struct rules that distinguish them in the new theory.
Thus the current strategy is to first find out whether or
not there are such cases with respect to $\tau$ in $C$. This is
a straightforward and quick operation.

The other problem in forming interpretations is the
possibility of multiple interpretations. For example,
consider the interpretation of the environment $dkh$ for
$\tau_4$ given in Figure 2, viz., $k_{\tau_6}$ ($dh$). If this inter-
pretation is adopted the original label for $\tau_3$, which was
$\delta \vee h$, will be changed to $dh$. (Whenever we specialize
a label by adding more specific environments to it, we
must remove any more general environments from the
label.) This is, in fact, the route that would be taken
by the current strategy. But another interpretation of
$dkh$ is possible, viz., $dhk \rightarrow \tau_4$ could be adopted as a
rule for $\tau_4$, and no changes would be made to the label
for $\tau_3$. Note, however, that while $h$ is rule-correlated to
$\tau_4$, $d$ is only theory-correlated to $\tau_4$. Adopting this
interpretation, therefore, has the effect of "promoting"
d to a rule-correlated-observable (rc-observable) of $\tau_4$.
In general, whenever possible, the current strategy fa-
vors interpretations that do not require such changes in
the status of observables or t-terms relative to the
$t$-term being retranslated.

**Figure 2**

<table>
<thead>
<tr>
<th>Environment</th>
<th>Interpretation</th>
<th>Modification</th>
</tr>
</thead>
<tbody>
<tr>
<td>$abeln$</td>
<td>$n_{\tau_3}$ ($abel$)</td>
<td>specialize $ael$ in label($\tau_3$)</td>
</tr>
<tr>
<td>$abdln$</td>
<td>$n_{\tau_3}$ ($abdl$)</td>
<td>specialize $adl$ in label($\tau_3$)</td>
</tr>
<tr>
<td>$bcgn$</td>
<td>$n_{\tau_3}$ ($bcg$)</td>
<td>generalize $bcfg$ in label($\tau_3$)</td>
</tr>
</tbody>
</table>

**Retranslation of $\tau_3$**

<table>
<thead>
<tr>
<th>Environment</th>
<th>Interpretation</th>
<th>Modification</th>
</tr>
</thead>
<tbody>
<tr>
<td>$abdl$</td>
<td>$b l_{\tau_2}$ ($ad$)</td>
<td>make $b$ rc-observable of $\tau_3$</td>
</tr>
<tr>
<td>$bdel$</td>
<td>$b l_{\tau_2}$ ($de$)</td>
<td>same</td>
</tr>
<tr>
<td>$abf$</td>
<td>$f_{\tau_1}$ ($ab$)</td>
<td>delete $g$ in rule $fg\tau_1 \rightarrow \tau_3$</td>
</tr>
<tr>
<td>$acfg$</td>
<td>$fg_{\tau_1}$ ($ac$)</td>
<td>-</td>
</tr>
<tr>
<td>$bcg$</td>
<td>$g_{\tau_1}$ ($hc$)</td>
<td>delete $f$ in rule $fg\tau_1 \rightarrow \tau_3$</td>
</tr>
</tbody>
</table>

**Environment Interpretation Modification**

| $abcn$ | $v_{\tau_3}$ ($abcn$) | specialize $abc$ in label($\tau_3$) |
| $bcfn$ | $v_{\tau_3}$ ($bcfn$) | specialize $bcv$ in label($\tau_3$) |

**New Theory**

| $ab \rightarrow \tau_1$, $da \rightarrow \tau_2$, $a \rightarrow \tau_7$ |
| $bl_{\tau_2} \rightarrow \tau_3$, $f_{\tau_1} \rightarrow \tau_3$, $g_{\tau_1} \rightarrow \tau_3$ |
| $n_{\tau_7} \rightarrow \tau_5$, $dh \rightarrow \tau_6$, $k_{\tau_6} \rightarrow \tau_4$ |
| $ach \rightarrow \tau_4$, $x_{\tau_7} \rightarrow \tau_8$, $ey_{\tau_7} \rightarrow \tau_8$ |

**Testing & Patching Interpretations**

Interpretations that involve the generalization or spe-
cialization of some label need to be tested against the
set of cases $C$. To see why, consider the example in
Figure 2, beginning with the retranslation of endpoint
$\tau_6$. In this case the interpretations of the environments
in RTLS-label($\tau_6$) lead to a $\delta$-label of $c$ for $\tau_7$. We see,
However, that if this interpretation of \( \tau_7 \) were adopted - other things being equal - a new false positive would be generated, i.e., the new label generated for \( \tau_8 \) would contain the environment \( cy \), where there are known cases containing \( cy \) that are non-\( \tau_7 \)-cases.

There are a number of options that can be pursued here. One that has proven to be useful, involves patching the interpretation \( y_7 \), i.e., adding more observables to it - so that the false positive will be avoided. Any observable that is not present in a problematic case but is present in every \( \tau_7 \)-case that \( cy \) is satisfied in, is a good candidate for a patch. Of course we prefer candidates that are either rule or theory correlated to \( \tau_8 \) in that order. In the example \( e \) fulfills this role, and leads to the adoption of the rule \( eyr_7 \rightarrow \tau_8 \). This and other patching techniques are similar to those discussed in [4].

Empirical Evaluation

As was mentioned above, top-down retranslation is a method for relaxed retranslation. This means that the new theory generated by this method may, and generally will, correspond to a reduction that is not identical to the input from RTLS. Therefore, it is conceivable that the error rate of the new theory - defined in terms of performance over all cases in the domain, and not just \( C \) - may be worse than that of the RTLS reduction.

While one would like to be able to say that a severe performance degradation cannot take place using this method, this remains unproven. However, experiments show that, if anything, one can expect top-down retranslation to lead to a new theory that gives better performance than the RTLS reduction. The evidence for this follows.

The top-down retranslation method described here has been tested on the same rheumatology knowledge base using the same 121 cases that were used to test RTLS [4]. As in the original testing of RTLS, the so-called leave-one-out method [6] for establishing an estimated error rate was employed. Using this method on \( n \) cases entails performing \( n \) trials over \( n - 1 \) of the cases, “leaving out” a different case each time to be used in testing the result of that trial. The estimated error is calculated by summing the errors over the \( n \) trials. Thus 121 trials were run, on each trial one case was set aside. RTLS was then run on the remaining 120 cases, and then top-down retranslation was applied to the RTLS reduction. The new theory was then tested on the case that was left out. An estimated error rate of 0 was obtained (RTLS achieved a .067 estimated error).

There is another dimension of performance along which a retranslation method must be tested. This is what we may call the “compression ratio.” This relates to the one of the avowed goals of retranslation, viz., to convert a rather large and cumbersome reduction into a smaller, more elegant, and more intelligible logical structure. In this case it is clear that the theory generated by top-down retranslation can be no worse than the RTLS reduction, the question is how much better is it likely to be?

Again, empirical results seem very reasonable. The rheumatology knowledge base initially consisted of roughly 300 rules which yielded a reduction of about 35,000 environments. The average size of the reduction produced by RTLS in the above experiments was roughly 30,000 environments, and the average number of rules generated by top-down retranslation was roughly 600. Technically, one ought to re-reduce the new theories in order to verify that they do indeed encode reductions on the order of 30,000 environments. In the interests of time, this was not done (it would probably take 10 or more hours to calculate each reduction), but cursory examination of the theories generated make it highly probable that this was in fact the case.

Conclusion

The results reported here show that the three-fold theory reduction approach to theory revision is feasible and robust. One would like to see the method tailored to work with partial reductions of theories, i.e., we want to reduce as little of the theory as possible to solve the revision problems at hand. This work establishes a framework and foundation within which such variations of top-down retranslation can be pursued.

References