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 $\mathcal{T}$ when $\mathcal{T}$ 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 $\mathcal{T}'$ of $\mathcal{T}$ which handles the set of all known cases correctly, makes use of the theoretical terms used in $\mathcal{T}$, 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, $\mathcal{C}$, but in such a way that it is reasonable to expect improved performance over cases not included in $\mathcal{C}$ 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 $\mathcal{C}$, 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'etre* 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).
τ-case: a c whose Answer(c) includes τ. **Non-τ-case:** a c whose Answer(c) does not include τ.
Rules-for(τ): the set of rules in theory $\mathcal{T}$ that directly conclude τ. **Level(τ):** max of the levels of Rules-for(τ)
δ-label(τ): a set of minimal environments being considered as a potential revised label for τ.
Endpoint: a t-term that does not occur in the antecedent of any rule.
RTLS-label(τ) the label generated for endpoint τ 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 δ-label
Rule-correlated-t-terms(τ): the set of all t-terms occurring in some member of Rules-for(τ)
Rule-correlated-observables(τ): the set of all observables occurring in some member of Rules-for(τ)
Theory-correlated-t-terms(τ): 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(τ) as the last link
Theory-correlated-observables(τ): the set of all observables that occur in any rule that is a "link" in a "rule-chain" having some rule in Rules-for(τ) as the last link

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

Let $\mathcal{T}$ be the theory and let the vocabulary (predicate symbols, propositional constants, etc.) of $\mathcal{T}$ be divided into two disjoint subsets $\mathcal{T}_o$ and $\mathcal{T}_t$. We refer to these as the *observational* (operational) and *theoretical* (non-operational) vocabulary of $\mathcal{T}$, respectively [8]. Let τ be a member of $\mathcal{T}_t$, and let $o_1 \ldots o_k$ be a conjunction of distinct items where $o_i \in \mathcal{T}_o$ for $i = 1, \ldots, k$. Suppose that the statement $o_1 \ldots o_k \rightarrow \tau$ follows from $\mathcal{T}$. 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 τ relative to $\mathcal{T}$. Now let $\mathcal{O}_\tau$ 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 τ relative to $\mathcal{T}$. Then $\mathcal{O}_\tau$ is the called the *reduction* of τ with respect to $\mathcal{T}_o$. Following the terminology of de Kleer [1], we sometimes call $\mathcal{O}_\tau$ the *label* for τ, and each member of $\mathcal{O}_\tau$ is said to be an *environment* for τ. The set of all $\mathcal{O}_\tau$ for $\tau \in \mathcal{T}_t$, denoted by $\mathcal{R}(\mathcal{T})$, is called the reduction of the theory $\mathcal{T}$.

Reduction of Expert System Theories

We consider an *expert system theory* $\mathcal{E}$ to be a *restricted* propositional logic theory. That is, $\mathcal{E}$ 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 $\mathcal{E}$ iff, to put it loosely, β can be derived from α and $\mathcal{E}$ via a sequence of applications of a generalized version of modus ponens. $\mathcal{E}$ is said to be *acyclic* if, roughly speaking, a sentence of the form $\alpha \rightarrow \alpha$ does not follow from $\mathcal{E}$.

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 $\mathcal{E}$ are partitioned into disjoint sets called *rule levels*. A rule r is in level 0 iff the truth-value of the left-hand side of r is a function of the truth-values of observables only. A rule r is in level n, iff the truth-value of the left-hand side of r is a function of the truth-values of observables and theoretical terms that are concluded only by rules at levels $0, \ldots, n-1$. 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 '*re*translation' 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 suggest that Newton had, in any sense, the *entire* reduction of his theory (or a variant thereof) available to him prior to its formulation in theoretical terms. Newton's laws entailed empirical generalizations that were not predicted and verified until well *after* the formulation 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 entirely 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 $\mathcal{R}(\mathcal{T}')$ for a theory $\mathcal{T}'$ 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 $\mathcal{R}(\mathcal{T}')$. Instead of viewing $\mathcal{R}(\mathcal{T}')$ 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 retranslation*. It is this version of retranslation that is most similar to the Newton-Kepler example. All that is required of the generated retranslation is that its performance over the cases $\mathcal{C}$ be at least as good as that of $\mathcal{R}(\mathcal{T}')$. In the sequel it is this version of the retranslation 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 *invented* 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 *structural* relationships among the various components of the theory - as embodied in its rules - provides crucial information in helping to guide the search for suitable retranslations. Recognizing the need {utility} for {of} new theoretical terms, while a relevant avenue for future investigation, is a task that is not directly addressed 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 reference, most of the special vocabulary used in this paper 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 $\mathcal{R}$ be the reduction we wish to retranslate, and let $\mathcal{T}$ be the version of the theory we were given prior to the learning session. For every *endpoint* $\tau \in \mathcal{T}_t$ - where τ is an endpoint iff it does not occur in the left-hand-side of any rule in $\mathcal{T}$ - a corresponding *RTLS label*, RTLS-label(τ), will exist in $\mathcal{R}$ (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, τ, we first try to find changes in the Rules-for(τ) and in the *labels* of the theoretical terms, *t-terms* for short, in these rules so that the label for τ generated by these changed rules and labels is either identical to, or fairly close to, RTLS-label(τ). What is important is that this label generated for τ - we call it a δ-label - yields the same performance results over the cases as RTLS-label(τ).

Intuitively this process corresponds to asking the question: What would the labels of the t-terms used to conclude τ - given by Rule-correlated-t-terms(τ) - as well as the new Rules-for(τ) have to "look like," in order for RTLS-label(τ) (or something "close enough" to it) to be the label that the retranslated theory will generate for τ? 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 τ, down one level of the theory. Let λ be any member of Rule-correlated-t-terms(τ). Now the question is: What would the labels of the t-terms in Rule-correlated-t-terms(λ) and the new Rules-for(λ) have to look like in order for the δ-label of λ to be generated?, and clearly we have to ask this question for every $\lambda \in$ Rule-correlated-t-terms(τ). 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 τ is a t-term at this level the new Rules-for(τ) will come directly from the δ-label(τ) 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

Figure 1

Original Theory $\mathcal{T}$

$ab \vee ac \vee bc \rightarrow \tau_1$, $ad \vee ae \vee ed \rightarrow \tau_2$

$fg\tau_1 \rightarrow \tau_3$, $h\tau_2 \rightarrow \tau_4$, $l\tau_2 \rightarrow \tau_3$, $n\tau_3 \rightarrow \tau_5$

$d \vee h \rightarrow \tau_6$, $k\tau_6 \rightarrow \tau_4$, $ce \rightarrow \tau_7$, $x\tau_7 \vee y\tau_7 \rightarrow \tau_8$

Endpoints of $\mathcal{T}$: τ_4, τ_5, τ_8

RTLS-label(τ_4): $adeh \vee dhk$

RTLS-label(τ_5): $abeln \vee abdln \vee bdeln \vee abfn \vee bcgn$

RTLS-label(τ_8): $cx \vee cey$

T-term	Level	Label in $\mathcal{T}$	Rule Correlated observables	Rule Correlated t-terms	Theory Correlated observables	Theory Correlated t-terms	Eigen-Terms
τ_1	0	$ab \vee ac \vee bc$	a, b, c	-	a, b, c	-	-
τ_2	0	$ad \vee ae \vee de$	a, d, e	-	a, d, e	-	-
τ_3	1	$abfg \vee acfg \vee bcfg \vee adl \vee ael \vee edl$	f, g, l	τ_1, τ_2	a, b, c, d, e, f, g, l	τ_1, τ_2	τ_1
τ_4	1	$adh \vee aeh \vee deh \vee dhk$	h, k	τ_2, τ_6	a, d, e, h, k	τ_2, τ_6	τ_6
τ_5	2	$abfgn \vee acfgn \vee bcfgn \vee adln \vee aeln \vee deln$	n	τ_3	$a, b, c, d, e, f, g, h, l$	τ_1, τ_2, τ_3	τ_1, τ_3
τ_6	0	$d \vee h$	d, h	-	d, h	-	-
τ_7	0	ce	c, e	-	c, e	-	-
τ_8	1	$cex \vee cey$	x, y	τ_7	c, e, x, y	τ_7	τ_7

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 $\mathcal{T}$, 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 Figure 1, we see that τ_2 is correlated to both endpoint τ_4 (a rule-correlation) and τ_5 (a theory-correlation). If the retranslation of τ_4 leads to a change in label for τ_2 this means we have to redo the retranslation of τ_5, assuming we did τ_5 first. One way to avoid this problem is simply to avoid changing the labels of any t-terms that effect more than one endpoint. T-terms that are theoretically-correlated to a single endpoint are called *eigen-terms* of that endpoint. In Figure 1, for example, we see that τ_6 is an eigen-term of τ_4, and that τ_1 and τ_3 are eigen-terms of τ_5, and that τ_7 is an eigen-term of τ_8. (Analogously, τ_1 is an eigen-term of τ_3.) By changing the labels of eigen-terms only (and by making sure that they remain eigen-terms in the final retranslation) we guarantee that no malicious interactions can occur by dividing up the retranslation problem as we have described. While this strategy can never fail, it may sometimes succeed too well, i.e., we may end up with a retranslated theory that makes less use of such *non*-eigen-terms than seems warranted. Ideally, one would like to modify only eigen-terms whenever possible, but when this fails to achieve good results the modification of non-eigen-terms should be considered. How to do so is a problem for future investigation.

Forming Interpretations

Suppose that we are trying to retranslate some endpoint, or other t-term, τ. This means that we have a δ-label(τ) at this point (either RTLS-label(τ) if τ is an endpoint, or else the current δ-label for τ as determined by the retranslation of the endpoint(s) to which τ is theoretically-correlated). We begin by identifying *Rule-correlated-observables*(τ) and *Rule-correlated-t-terms*(τ), where these are the observables and t-terms that occur in some rule that directly concludes τ. We now try to "interpret" or "reconstruct" δ-label(τ) by finding a set of rules for τ 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 τ_5 can be viewed as arising from the rule $n\tau_3 \rightarrow \tau_5$ provided that the appropriate modifications to the label of τ_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 τ_3.

There are three activities included in the interpretation-forming phase. In the first place we are generating candidates for the new Rules-for(τ). 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 δ-labels of the t-terms that are

used to conclude τ. Consider, for example, the retranslation of τ_5 in Figure 2. In this case each desired environment happens to generate the same interpretation $n\tau_3$ (which is, in fact, identical to a rule in the original theory), but each environment "impacts" a different environment from the original label of τ_3. For example, the desired environment *abeln* forces a *specialization* of the environment *ael* in the original label to the environment *abel* in the new label, while the environment *bcgn* forces a *generalization* of the environment *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, τ_3 in the example, accurately reflects *all* the changes arising from the interpretations that are, at least tentatively, being considered. In the current system this is achieved by obeying the following regimen. We first perform all the specialization modifications to the original 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 perform 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 impossible to interpret all the environments of δ-label(τ) in terms of the items in *Rule-correlated-observables*(τ) and *Rule-correlated-t-terms*(τ). This will certainly be the case if some $e \in \delta$-label(τ) contains one or more observables that are not in *theory-correlated-observables*(τ).

In fact it is easy to know in advance whether or not *theory-correlated-observables*(τ) will have to be augmented 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 observables for* τ, then we know that we will have to make use of the other observables in these cases if we are to construct 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 τ in $\mathcal{C}$. 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 *dhk* for τ_4 given in Figure 2, viz., $k\tau_6$ (*dh*). If this interpretation is adopted the original label for τ_6, which was $d \vee h$, 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 *dhk* is possible, viz., $dhk \rightarrow \tau_4$ could be adopted as a rule for τ_4, and no changes would be made to the label for τ_6. Note, however, that while h is rule-correlated to τ_4, d is only theory-correlated to τ_4. Adopting this interpretation, therefore, has the effect of "promoting" d to a rule-correlated-observable (rc-observable) of τ_4. In general, whenever possible, the current strategy favors interpretations that do not require such changes in the status of observables or t-terms relative to the t-term being retranslated.

Figure 2

Retranslation of τ_8

Environment	Interpretation	Modification
cx	$x\ \tau_7$ (**c**)	generalize *ce* in label(τ_7)
cey	$y\ \tau_7$ (**ce**)	-

δ-label for τ_7: *c*

Resulting label(τ_8): $cx \vee cy$, but *cy* leads to false positives for τ_8

Patch: Add e to problematic interpretation, i.e., rule $y\tau_7 \rightarrow \tau_8$ becomes $ey\tau_7 \rightarrow \tau_8$

Retranslation of τ_5

Environment	Interpretation	Modification
abeln	$n\ \tau_3$ (**abel**)	specialize *ael* in label(τ_3)
abdln	$n\ \tau_3$ (**abdl**)	specialize *adl* in label(τ_3)
bdeln	$n\ \tau_3$ (**bdel**)	specialize *del* in label(τ_3)
abfn	$n\ \tau_3$ (**abf**)	generalize *abfg* in label(τ_3)
bcgn	$n\ \tau_3$ (**bcg**)	generalize *bcfg* in label(τ_3)

δ-label for τ_3: $abdl \vee abel \vee bdel \vee abf \vee acfg \vee bcg$

Retranslation of τ_3

Environment	Interpretation	Modification
abdl	$b\ l\ \tau_2$ (**ad**)	make *b* rc-observable of τ_3
abel	$b\ l\ \tau_2$ (**ae**)	*same*
bdel	$b\ l\ \tau_2$ (**de**)	*same*
abf	$f\ \tau_1$ (**ab**)	delete *g* in rule $fg\tau_1 \rightarrow \tau_3$
acfg	$fg\ \tau_1$ (**ac**)	-
bcg	$g\ \tau_1$ (**bc**)	delete *f* in rule $fg\tau_1 \rightarrow \tau_3$

δ-labels for τ_1 & τ_2: identical to their original labels

Resulting label(τ_3): $abdl \vee abel \vee bdel \vee abf \vee acf \vee bcf \vee abg \vee acg \vee bcg$

Resulting label(τ_5): $abeln \vee abdln \vee bdeln \vee abfn \vee acfn \vee bcfn \vee abgn \vee acgn \vee bcgn$

Retranslation of τ_4

Environment	Interpretation	Modification
adeh	*adeh*	Add rule: $adeh \rightarrow \tau_4$
dhk	$k\ \tau_6$ (**dh**)	specialize *dh*

New Theory

$ab \vee ac \vee bc \rightarrow \tau_1$, $ad \vee ae \vee ed \rightarrow \tau_2$, $c \rightarrow \tau_7$
$bl\tau_2 \rightarrow \tau_3$, $f\tau_1 \rightarrow \tau_3$, $g\tau_1 \rightarrow \tau_3$
$n\tau_3 \rightarrow \tau_5$, $dh \rightarrow \tau_6$, $k\tau_6 \rightarrow \tau_4$
$adeh \rightarrow \tau_4$, $x\tau_7 \rightarrow \tau_8$, $ey\tau_7 \rightarrow \tau_8$

Testing & Patching Interpretations

Interpretations that involve the generalization or specialization of some label need to be tested against the set of cases $\mathcal{C}$. To see why, consider the example in Figure 2, beginning with the retranslation of endpoint τ_8. In this case the interpretations of the environments in RTLS-label(τ_8) lead to a δ-label of c for τ_7. We see,

however, that if this interpretation of τ_7 were adopted - other things being equal - a new false positive would be generated, i.e., the new label generated for τ_8 would contain the environment *cy*, where there are known cases containing *cy* that are *non*-τ_7-cases.

There are a number of options that can be pursued here. One that has proven to be useful, involves *patching* the interpretation $y\tau_7$, 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 τ_8-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 τ_8 in that order. In the example *e* fulfills this role, and leads to the adoption of the rule $ey\tau_7 \rightarrow \tau_8$. 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 360 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

[1] J. de Kleer. An assumption-based tms. *Artificial Intelligence*, 28:127–162, 1986.

[2] A. Ginsberg. Knowledge base refinement and theory revision. In *Proceedings of The Sixth International Workshop on Machine Learning*, pages 260–265, 1989.

[3] A. Ginsberg. Knowledge-base reduction: a new approach to checking knowledge bases for inconsistency and redundancy. In *Proceedings of the Seventh Annual National Conference on Artificial Intelligence*, pages 585–589, 1988.

[4] A. Ginsberg. Theory revision via prior operationalization. In *Proceedings of the Seventh Annual National Conference on Artificial Intelligence*, pages 590–595, 1988.

[5] C. Hempel. *Philosophy of Natural Science.* Prentice-Hall, Englewood Cliffs, N.J., 1966.

[6] P. Lachenbruch. An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics*, 24:639–645, December 1967.

[7] T. Mitchell, R. Keller, and S. Kedar-Cabelli. Explanation-based generalization: a unifying view. *Machine Learning*, 1:47–80, 1986.

[8] E. Nagel. *The Structure of Science.* Harcourt, Brace, and World, New York, 1961.