VALIDATION OF TERMINOLOGICAL INFERENCE IN AN INFORMATION EXTRACTION TASK

Marc Vilain

The MITRE Corporation
Burlington Rd.
Bedford, MA 01730
mbv@linus.mitre.org

ABSTRACT

This paper is concerned with an inferential approach to information extraction, reporting in particular on the results of an empirical study that was performed to validate the approach. The study brings together two lines of research: (1) the RHO framework for tractable terminological knowledge representation, and (2) the *Alembic* message understanding system. There are correspondingly two principal aspects of interest to this work. From the knowledge representation perspective, the present study serves to validate experimentally a normal form hypothesis that guarantees tractability of inference in the RHO framework. From the message processing perspective, this study substantiates the utility of limited inference to the information extraction task.

1. SOME BACKGROUND

Alembic is a natural language-based information extraction system that has been under development for about one year. As with many such systems, the information extraction process in *Alembic* occurs through pattern matching against the semantic representation of sentences. These representations are themselves derived from parsing the input text, in our case with a highly lexicalized neo-categorial grammar [1].

Experience has shown that this kind of approach can yield impressive performance levels in the data extraction task (see [18]). We have found—as have others—that meaningful results can be obtained despite only having sketchy sentence semantics (as can happen when there are widespread gaps in the lexicon's semantic assignments). In addition, because the parsing process normalizes the sentence semantics to a significant degree, the number of extraction patterns can be relatively small, especially compared to approaches that use only rudimentary parsing.

Strict semantic pattern-matching is unattractive, however, in cases that presume some degree of inference. Consider the following example of an East-West joint venture:

> [...] Samsung signed an agreement with Soyuz, the external-trade organization of the Soviet Union, to swap Korean TV's and VCR's for pig iron from the Soviet Union

What makes this sentence an example of the given joint venture concept is an accumulation of small inferences: that Soyuz is a Soviet entity, that signing an agreement designates agreement between the signing parties, and that the resulting agreement holds between a Soviet and non-Soviet entity. Such examples suggest that it is far preferable to approach the extraction problem through a set of small inferences, rather than through some monolithic extraction pattern. This notion has been embodied in a number of earlier approaches, e.g. [11] or [17].

The inferential approach we were interested in bringing to bear on this problem is the RHO framework. RHO is a terminological classification framework that ultimately descends from KL-ONE. Unlike most recent such systems, however, RHO focuses on terminological inference (rather than subsumption). And whereas most KL-ONE descendants sacrifice completeness for computational tractability, inference in RHO is complete in polynomial time if terminological axioms meet a normal form criterion.

Nevertheless, before embarking on a significant development effort to implement the RHO framework under *Alembic*, we wanted to verify that the framework was up to the data extraction task. In particular, we were keen to ensure that the theoretical criterion that guarantees polynomial time completeness for RHO was actually met in practice. Towards this end, my colleagues and I undertook an extensive empirical study whose goal was, among others, to validate this criterion.

The present paper is a summary of our findings, with a special focus on RHO itself and on the validation task. We provide some suggestive interpretations of these findings, and touch on current and ongoing work towards bringing RHO to bear on the extraction task in *Alembic*.

2. THE RHO FRAMEWORK

The RHO framework, as noted above, arose in reaction to standard approaches to terminological reasoning, as embodied in most descendants of KL-ONE, e.g., CLASSIC [4], BACK [13], LOOM [12], and many others. This line of work has come to place a major emphasis on computing concept

subsumption, i.e., the determination of whether a representational description (a concept) necessarily entails another description. In our view, this emphasis is mistaken.

Indeed, this emphasis ignores the way in which practical applications have successfully exploited the terminological framework. These systems primarily rely on the operation of classification, especially instance classification. Although subsumption helps to provide a semantic model of classification, it does not necessarily follow that it should provide its computational underpinnings.

In addition, the emphasis on complete subsumption algorithms has led to restricted languages that are representationally weak. As is well-known, these languages have been the subject of increasingly pessimistic theoretical results, from intractability of subsumption [5], to undecidability of subsumption [15, 16], to intractability of the fundamental normalization of a terminological KB [14].

Against this background, RHO was targeted to support instance classification, and thus departs in significant ways from traditional terminological reasoners. The most draconian departure is in separating the normal terminological notion of necessary and sufficient definitions into separate sufficiency axioms and necessity axioms. The thrust of the former is to provide the kind of antecedent inference that is the hallmark of classification, e.g.,

$$\text{western-corp}(x) \leftarrow \text{corporation}(x) \,\&\, \text{hq-in}(x, y) \,\&\, \text{western-nation}(y) \quad (1)$$

The role of necessity conditions is to provide consequent inference such as that typically associated with inheritance and sort restrictions on predicates, e.g.,

$$\text{organization}(x) \leftarrow \text{corporation}(x) \quad (2)$$
$$\text{corporation}(x) \leftarrow \text{western-corp}(x) \quad (3)$$
$$\text{organization}(x) \leftarrow \text{agreement}(x, y, z) \quad (4)$$

Although both classes of axioms are expressed in the same syntactic garb, namely function-free Horn clauses, they differ with respect to their inferential import. If one thinks of predicates as being organized according to some taxonomy (see Fig. 1), then necessity axioms encode inference that proceeds up the hierarchy (i.e., inheritance), while sufficiency axioms encode inference that proceeds down the hierarchy (i.e., classification).

The most interesting consequence of RHO's uniform language for necessity and sufficiency is that it facilitates the formulation of a criterion under which classification is guaranteed to be tractable. For a knowledge base to be guaranteed tractable, the criterion requires that there be a tree shape to the implicit dependencies between the variables in any given axiom in the knowledge base.

For the sample axioms above, Fig. 2 informally illustrates this notion of variable dependencies. Axiom (1), for example, mentions two variables, x and y. A dependency between these variables is introduced by the predicative term hq-in(x,y): the term makes the two variables dependent by virtue of mentioning them as arguments of the same predicate. As the axiom mentions no other variables, its dependency graph is the simple tree on the left of Fig. 1. Similarly, in axiom (4) the *agreement* predicate makes both y and z dependent on x, also yielding a tree. Finally, axioms (2) and (3) lead to degenerate trees containing only x. Since all the dependency relations between these variables are tree-shaped, the knowledge base formed out of their respective axioms is tractable under the criterion. A formal proof that tractability follows from the criterion appears in an appendix below, as well as in [19].

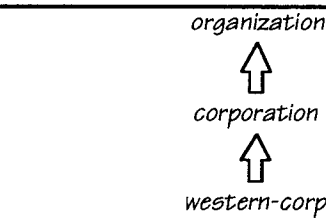

Figure 1: A predicate taxonomy

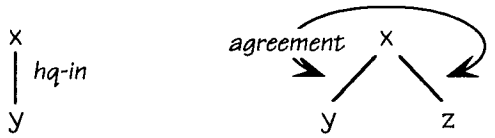

Figure 2: Dependency trees for variables in axioms (1), on the left, and (4), on the right.

3. VALIDATING RHO

This formal tractability result is appealing, especially in light of the overwhelming number of intractability claims that are usually associated with terminological reasoning. Its correctness, however, is crucially dependent on a normal form assumption, and as with all such normal form criteria, it remains of little more than theoretical interest unless it is validated in practice. As we mentioned above, we strove to achieve such a validation by determining through a paper study whether the RHO framework could be put to use in the data extraction phase of *Alembic*.

Towards this end, my colleagues and I assembled a set of unbiased texts on Soviet economics. The validation task then consisted of deriving a set of terminological rules that would allow RHO to perform the inferential pattern matching necessary to extract from these texts all instances of a pre-determined class of target concepts. The hypothesis that RHO's tractability criterion can be met in practice would thus be considered validated just in case this set of inference rules was tractable under the criterion.

3.1. Some assumptions

At the time that we undertook the study, however, the *Alembic* implementation was still in its infancy. We thus had to make a number of assumptions about what could be expected out of *Alembic*'s parsing and semantic composition components. In so doing, we took great pain not to require superhuman performance on the part of the parser, and restricted our expected syntactic coverage to phenomena that we felt were well within the state of the art.

In particular, we rejected the need to derive S. As with many similar systems, *Alembic* uses a fragment parser that produces partial syntactic analyses when its grammar is insufficient to derive S. In addition, we exploited *Alembic*'s hierarchy of syntactic categories, and postulated a number of relatively fine-grained categories that were not currently in the system. This allowed us for example to assume we could obtain the intended parse of "Irish-Soviet airline" on the basis of the pre-modifiers being both adjectives of geographic origin (and hence co-ordinable).

We also exploited the fact that the *Alembic* grammar is highly lexicalized (being based on the combinatorial categorial framework). This allowed us to postulate some fairly detailed subcategorization frames for verbs and their nominalizations. As is currently the case with our system, we assumed that verbs and their nominalizations are canonicalized to identical semantic representations.

Elsewhere at the semantic level, we assumed basic competence at argument-passing, a characteristic already in place in the system. This allowed us, for example, to assume congruent semantics for the phrases "Samsung was announced to have X'd" and "Samsung has X'd."

3.2. The validation corpus

With these assumptions in mind, we assembled a corpus of data extraction inference problems in the area of Soviet economics. The corpus consisted of text passages that had been previously identified for an evaluation of information retrieval techniques in this subject area. The texts were drawn from over 6200 Wall Street Journal articles from 1989 that were released through the ACL-DCI. These articles were filtered (by extensive use of GREP) to a subset of 300-odd articles mentioning the then-extant Soviet Union. These articles were read in detail to locate all passages on a set of three pre-determined economic topics:

1. East-West joint ventures, these being any business arrangements between Soviet and non-Soviet agents.
2. Hard currency, being any discussion of attempts to introduce a convertible unit of monetary value in the former USSR.
3. Private cooperatives, i.e., employee-owned enterprises within the USSR.

We found 85 such passages in 74 separate articles (1.2% of the initial set of articles under consideration).

Among these, 47 passages were eliminated from consideration because they were just textual mentions of the target concepts (e.g. the string "joint venture") or of some simple variant. These passages could easily be identified by Boolean keyword techniques, and as such were not taken to provide a particularly insightful validation of a complex NL-based information-extraction process! Unfortunately, this eliminated all instances of private cooperatives from the corpus, because in these texts, the word "cooperative" is a perfect predictor of the concept.

An additional four passages were also removed during a cross-rater reliability verification. These were all amplifications of an earlier instance of one of the target concepts, e.g., "U.S. and Soviet officials hailed the joint project." These passages were eliminated because the corpus collectors had differing intuitions as to whether they were sufficient indications in and of themselves of the target concepts, or were somehow pragmatically "parasitic" upon earlier instances of the target concept. The remaining 34 passages required some degree of terminological inference, and formed the corpus for this study.

4. INFERENTIAL DATA EXTRACTION

We then set about writing a collection of terminological axioms to handle this corpus. As these axioms are propositional in nature, and the semantic representations produced by *Alembic* are not strictly propositional, this required specifying a mapping from the language of interpretations to that of the inference axioms.

4.1. Semantic representation in *Alembic*

Alembic produces semantic representations at the increasingly popular interpretation level [2, 10]. That is, instead of generating fully scoped and disambiguated logical forms, *Alembic* produces representations that are ambiguous with respect to quantifier scoping. For example, the noun phrase "a gold-based ruble" maps into something akin to the following interpretation:

```
[ [head ruble]
  [quant :exists]
  [args NIL]
  [proxy P117]
  [mods { [ [head basis-of]
            [args { P117 [ [head gold]
                           [quant :kind]] }]]}]]
```

Semantic heads of phrases are mapped to the *head* slot of the interpretation, arguments are mapped to the *args* slot,

modifiers to the mods slot, and generalized quantifiers to the quant slot. The proxy slot contains a unique variable designating the individuals that satisfy the interpretation. If this interpretation were to be fully mapped to a sorted first-order logical form, it would result in the following sentence, where *gold* is treated as a kind individual:

$\exists$ P117 : ruble basis-of(P117, *gold*)

Details of this semantic framework can be found in [3].

4.2 Conversion to propositional form

Axioms in RHO are strictly function-free Horn clauses, and as such are intended to match neither interpretations nor first-order logical forms. As a result, we needed to specify a mapping from interpretations to some propositional encoding that can be exploited by RHO's terminological axioms. In brief, this mapping hyper-Skolemizes the proxy variables in the interpretation and then recursively flattens the interpretation's modifiers.[1]

For example, the interpretation for "a gold-based ruble" is mapped to the following propositions:

ruble(P117)
basis-of(P117, *gold*)

The interpretation has been flattened by pulling its modifier to the same level as the head proposition (yielding an implicit overall conjunction). In addition, the proxy variable has been interpreted as a Skolem constant, in this case the "gensymed" individual P117.

This interpretation of proxies as Skolem constants is actually hyper-Skolemization, because we perform it on universally quantified proxies as well as on existentially quantified ones. Ignoring issues of negation and disjunction, this unorthodox Skolemization process has a curious model-theoretic justification (which is beyond our present scope). Intuitively, however, one can think of these hyper-Skolemized variables as designating the individuals that would satisfy the interpretation, once it has been assigned some unambiguously scoped logical form.

To see this, say we had the following inference rule:

m-loves-w(x,y) $\leftarrow$ loves (x, y) & man (x) & woman(y)

Now say this rule were to be applied against the semantics of the infamously ambiguous case of "every man loves a woman." In propositionalized form, this would be:

man(P118)
woman(P119)
loves(P118,P119)

[1]This glosses over issues of event reference, which we address through a partly Davidsonian framework, as in [9].

target	*occurrences, n*	*sufficiency rules, r*	*rule density, r/n*
joint venture	12	17	1.4
hard curr.	22	13	.59

Table 1: Summary of experimental findings.

From this, the rule will infer m-loves-w(P118,P119). If we think of P118 and P119 as designating the individuals that satisfy the logical form of "every man loves a woman" in some model, then we can see that indeed the m-loves-w relation necessarily must hold between them. This is true regardless of whether the model itself satisfies the standard $\forall$-$\exists$ scoping of the sentence or the notorious $\exists$-$\forall$ scoping. This demonstrates a crucial property of this approach, namely that it enables inferential extraction over ambiguously scoped text, without requiring resolution of the scope ambiguity (and without expensive theorem proving).

5. FINDINGS

Returning to our validation study, we took this propositionalized representation as the basis for writing the set of axioms necessary to cover our corpus of data extraction problems. In complete honesty, we expected that the resulting axioms would not all end up meeting the tractability criterion. Natural language is notoriously complex, and even such classic simple KL-ONE concepts as Brachman's arch [6] do not meet the criterion.

What we found took us by surprise. We came across many examples that were challenging at various levels: complex syntactic phenomena, nightmares of reference resolution, and the ilk. However, once the corpus passages were mapped to their corresponding interpretations, the terminological axioms necessary to perform data extraction from these interpretations all met the criterion.

Table 1, above, summarizes these findings. To cover our corpus of 34 passages, we required between two and three dozen sufficiency rules, depending upon how one encoded certain economic concepts, and depending on what assumptions one made about argument-passing in syntax. We settled on a working set of thirty such rules.

Note that this inventory does not include any necessity rules. We ignored necessity rules for the present purposes in part because they only encode inheritance relationships. The size of their inventory thus only reflects the degree to which one chooses to model intermediate levels of the domain hierarchy. For this study, we could arguably have used none. In addition, necessity rules are guaranteed to meet the tractability criterion, and were consequently of only secondary interest to our present objectives.

5.1. Considerations for data extraction

From a data extraction perspective, these results are clearly preliminary. Looking at the positive side, we are encouraged that the rules for our hard currency examples were shared over multiple passages, as follows from their fractional rule density of .59 (see Table 1). The joint venture rules fared less well, mainly because the concept they encode is fairly complex, and can be described in many ways.

Given our restricted data set, however, it is not possible to conclude how well either set of rules will generalize if presented with a larger corpus. What is clearly needed is a larger corpus of examples. This would allow us to estimate generalizability of the rules by considering the asymptotic growth of the rule set as it is extended to cover more examples. Unfortunately, constructing such a corpus is a laborious task, since the examples we are interested in are precisely those that escape simple automated search techniques such as Boolean keyword patterns. The time and expense that were incurred in constructing the MUC3/4 and TIPSTER corpora attest to this difficulty.

We soon hope to know more about this question of rule generalizability. We are currently in the process of implementing a version of RHO in the context of the *Alembic* system, which is now considerably more mature than when we undertook the present study. We intend to exploit this framework for our participation in MUC5, as well as retool our system for the MUC4 task. As the TIPSTER and MUC4 data sets contain a considerably greater number of training examples than our Soviet economics corpus, we expect to gain much better insights into the ways in which our rule sets grow and generalize.

5.2. Considerations for RHO

From the perspective of our terminological inference framework, however, these preliminary results are quite encouraging indeed. We started with a very simple tractable inference framework, and studied how it could be applied to a very difficult problem in natural language processing. And it appears to work.

Once again, one should refrain from reaching overly general conclusions based on a small test sample. And admittedly RHO gets a lot of help from other parts of *Alembic*, especially the parser and a rudimentary inheritance taxonomy. Further analyses, however, reveal some additional findings that suggest that RHO's tractability criterion may be of general validity to this kind of natural language inference.

Most interestingly, the tractability result can be understood in the context of some basic characteristics of natural language sentence structure. In particular, axioms that violate the tractability criterion can only be satisfied by sentences that display anaphora or definite reference. For example, an axiom with the following right hand side:

$$own(x, z) \;\&\; scorn(x, y) \;\&\; dislike(y, z)$$

matches the sentences "the man who owns a Ferrari scorns anyone who dislikes it/his car/that car/the car." It is impossible, however, to satisfy this kind of circular axiom without invoking one of these referential mechanisms (at least in English). This observation, which was made in another context in [8], suggests a curious alignment between tractable cases of terminological natural language inference and non-anaphoric cases of language use.

It is particularly tantalizing that the cases where these terminological inferences are predicted to become computationally expensive are just those for which heuristic interpretation methods seem to play a large role (e.g., discourse structure and other reference resolution strategies). Though one must avoid the temptation to draw too strong a conclusion form such coincidences, one is still left thinking of Alice's ineffable words, "Curiouser and curiouser."

Acknowledgments

Much gratitude is owed John Aberdeen for preparing our corpus through tireless perusal of the Wall Street Journal. Many thanks also to those who served as technical inspiration or as sounding boards: Bill Woods, Remko Scha, Steve Minton, Dennis Connolly, and John Burger.

REFERENCES

[1] Aberdeen, J., Burger, J., Connolly, D., Roberts, S., & Vilain, M. (1992). "Mitre-Bedford: Description of the Alembic system as used for MUC-4". In [18].

[2] Alshawi, H. & Van Eijck, J. (1989). "Logical forms in the core language engine". In *Prcdgs. of ACL89*. Vancouver, BC, 25-32.

[3] Bayer, S. L. & Vilain, M. B. (1991). "The relation-based knowledge representation of King Kong". *Sigart Bulletin* 2(3), 15-21.

[4] Brachman, R. J., Borgida, A., McGuiness, D. L., & Patel-Schneider, P. F. (1991). "Living with CLASSIC". In Sowa, J., *ed.*, *Principles of Semantic Networks*. San Mateo, CA: Morgan-Kaufmann.

[5] Brachman, R. J. & Levesque, H. (1984). "The tractability of subsumption in frame-based description languages". In *Prcdgs. of AAAI84*. Austin, Texas, 34-37.

[6] Brachman, R. J. & Schmolze, J. K. (1985). "An overview of the KL-ONE knowledge representation system". *Cognitive Science* 9(2), 171-216.

[7] Garey, M. R. & Johnson, D. S. (1979). *Computers and Intractability*. New York: W. H. Freeman.

[8] Haddock, N. J., (1992). "Semantic evaluation as constraint network consistency". In *Prcdgs. of AAAI92*. San Jose, CA, 415-420.

[9] Hobbs, J. R. (1985). "Ontological promiscuity". In *Prcdgs. of ACL85*. Chicago, IL, 119–124.

[10] Hobbs, J. R. & Shieber, S. M. (1987). "An algorithm for generating quantifier scopings". *Computational Linguistics* 13(1-2), 47-63.

[11] Jacobs, P. S. (1988). "Concretion: Assumption-based understanding". In *Prcdgs. of the 1988 Intl. Conf. on Comput. Linguistics (COLING88)*. Budapest, 270-274.

[12] MacGregor, R. (1991). "Inside the LOOM description classifier". *Sigart Bulletin* 2(3), 88-92.

[13] Nebel, B. (1988). "Computational complexity of terminological reasoning in BACK". *Artificial Intelligence* 34(3), 371-383.

[14] Nebel, B. (1990). "Terminological reasoning in inherently intractable". *Artificial Intelligence* 43, 235-249.

[15] Patel-Schneider, P. F. (1989). Undecidability of subsumption in NIKL. *Artificial Intelligence* 39: 263–272.

[16] Schmidt-Schauß, M. (1989). "Subsumption in KL-ONE is undecidable". In *Prcdgs. of KR89*. Toronto, ON.

[17] Stallard, D. G. (1986). "A terminological simplification transformation for natural language question-answering systems". In *Prcdgs. of ACL86*. New York, NY, 241-246.

[18] Sundheim, B., *ed.* (1992). *Prcdgs. of the Fourth Message Understanding Conf.* (MUC-4), McLean, VA, 215-222.

[19] Vilain, M. (1991). "Deduction as parsing: Tractable classification in the KL-ONE framework". In *Prcdgs. of AAAI91*. Anaheim, CA, 464-470

APPENDIX: PROOF OF TRACTABILITY

To demonstrate the validity of the tractability criterion, we only need consider the computational cost of finding all instantiations of the right-hand side of an axiom. In general, finding a single such instantiation is NP-complete, by reduction to the conjunctive Boolean query problem (see [7]). Intuitively, this is because general function-free Horn clauses can have arbitrary interactions between the variables on the right-hand side, i.e., their dependency graphs are fully cross-connected, as in:

$$R(v_1,v_2) \,\&\, R(v_1,v_3) \,\&\, R(v_2,v_3) \,\&\, R(v_1,v_4) \,\&\, R(v_2,v_4)\ldots$$

Intuitively again, verifying the instantiation of a given variable in a rule may require (in the worst case) checking all instantiations of all other variables in the rule. Under the usual assumptions of NP-completeness, no known algorithm exists that performs better in the worst case than enumerating all these instantiations. As each variable may take on as many as κ instantiations, where κ is the number of constants present in the knowledge base, the overall cost of finding a single globally consistent instantiation is $O(\kappa^{\xi})$, where ξ is the number of variables in the rule. The resulting complexity is thus exponential in ξ, which itself varies in the worst case with the length of the rule.

Consider now an axiom that satisfies the tractability criterion, yielding a graph such as that in Fig. 3. By

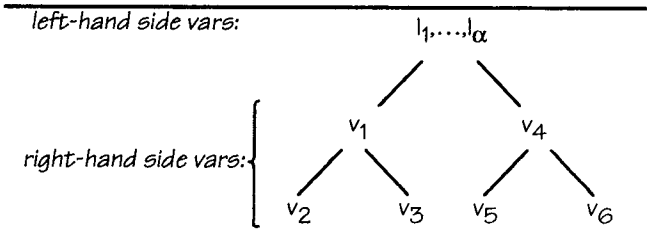

Figure 3: A dependency graph.

definition, the root of the graph corresponds to all the variables on the left-hand side, and all other nodes correspond to some variable introduced on the right-hand side. The cost of finding all the instantiations of the root variables is bounded by κ^{α}, where α is the maximal predicate valence for all the predicates appearing in the database. The cost of instantiating each non-root variable v is in turn bounded by $\alpha\kappa^{\alpha}$, corresponding to the cost of enumerating all possible instantiations of any predicate relating v to its single parent in the graph.

The topological restriction of the criterion leads directly to the fact that the exponent of these terms is a low-magnitude constant, α, rather than a parameter, ξ, that can be allowed to grow arbitrarily with the complexity of inference rules. The topological restriction also leads to the fact that these terms contribute *additively* to the overall cost of finding all instantiations of a rule. This overall cost is thus bounded by $\kappa^{\alpha} + \underbrace{\alpha\kappa^{\alpha}+\ldots+\alpha\kappa^{\alpha}}_{\xi}$, or $O(\xi\alpha\kappa^{\alpha})$.

Finally, we note that with the appropriate indexing scheme, finding all consequents of all rules only adds a multiplicative cost of ρ, where ρ is the total number of rules, yielding a final overall cost of $O(\rho\xi\alpha\kappa^{\alpha})$. It is often assumed that predicates in natural languages have no more than three arguments, so this formula approximately reduces to $O(\kappa^3)$.

This is of course a worst-case estimate. We are looking forward to measuring run-time performance figures on the MUC5 task, and are of course hoping to find actual performance to lie well below this cubic upper bound.