delim $$

Proving Self-Utterances

$Michael˜Miller sup a,b˜and˜˜Donald˜Perlis sup a,c$

MX$"" sup a Computer˜Science MZ"" sup b Systems˜Research MY"" sup c Institute˜for˜Advanced$
Department Center Computer Studies

University of Maryland
College Park, Maryland 20742
(301) 454-2002

Abstract

We study the *Knights and Knaves* problem, and find that for a proper treatment via theorem-proving, an interaction with natural language processing research is helpful. In particular, we discuss Ohlbach's claim that first-order logic is not well suited to handling this problem. Then we provide an interpretation of the problem using indexicals, axiomatize it, and prove the desired result. We conclude by suggesting a broader context for dealing with ''self-utterances'' in automatic theorem-proving.

descriptors: indexicals, knowledge representation, first-order logic, situated logic, utterances

I. Introduction

The *Knights and Knaves* problem [Smullyan 1978] can be stated as follows: An island exists whose only inhabitants are knights, knaves, and a princess. The knights on the island always tell the truth, while the knaves always lie. Some of the knights are poor and the rest of them are rich. The same holds for the knaves. The princess is looking for a husband who must be a rich knave. In uttering one statement, how can a rich knave convince the princess that he is indeed a prospective husband for her? [1]

[Ohlbach 1985] is devoted to the framing and solution of this problem in a formal theorem-proving context using first-order logic (FOL). Though trying to write the problem in FOL may not appear to be difficult at first, it is shown by Ohlbach not to be entirely elementary. He examines, and finds inadequate, two different approaches before he finally settles on a third. This final approach, though successful in that it gets the desired ''solution'', is unsatisfactory. Specifically, Ohlbach uses a truth predicate with two arguments, T(x,y), which claims that its first argument x is true, but has no clear meaning for its second

[1] The intended solution is the statement ''I am a poor knave.'' The reader can readily verify that this indeed is a solution. Note the self-referential nature of the statement; this feature is a special case of indexicality, which we address below. For general treatments of self-reference, see [Perlis 1985], where another of Smullyan's puzzles is treated, and [Smith 1986].

argument in terms of the original problem. Its justification is that the predicate allows a theorem-prover to perform certain unifications that lead to the intended solution. But it does not accomplish the goal of finding a knowledge representation faithful to the original problem, as well as having the solution as a logical consequence.

Ohlbach's conclusion is that knowledge representation is too hard in first-order logic, and too dependent upon tricks. We will not dispute that it takes some time to come up with a satisfactory representation of the problem, but this is not necessarily the fault of first-order logic. We contend that there is a straightforward treatment of the problem that is faithful to its intent and that does allow a formal proof of the desired result. However, it requires employing concepts into the formalism that are not usually found in the context of problem-solving via resolution theorem-provers, namely, ideas from natural language processing. Nevertheless, we are not replacing one trick by another, but rather introducing a well-understood and general formalism for problems of this sort.

The rest of this paper is organized as follows: section II discusses issues of problem representation, especially the role played by the pronoun ''I'' in the *Knights and Knaves* problem. Section III reviews general consequences for truth-values of statements containing indexicals such as ''I'', and section IV applies this to the *Knights and Knaves* problem. Section V gives our formal treatment, including a resolution-refutation proof by answer extraction. Section VI compares out solution to Ohlbach's and suggests a broader context for dealing with self-utterances in automatic theorem-proving.

II. Problem Representation

Finding a suitable representation for problems in artificial intelligence (AI) is often a difficult task. However, the *formalism* used to represent a problem is not necessarily the cause of the difficulty, though we grant that sometimes it is. Often it is the problem itself that is resisting representation and, when this occurs, further insight into the problem is necessary.

The *Knights and Knaves* problem is a prime example of this. Ohlbach's interpretation of the problem results in his asking ''Is there a statement x that I (being a rich knave) can say to convince the princess that I am indeed a rich knave?'' Formally this might look like (and does in Ohlbach's second treatment):

OHL: $(\exists u)[\mathrm{CanSay}(I,u) \leftrightarrow T(\mathrm{and}(\mathrm{knave}(I),\mathrm{rich}(I)))]$[2]

where T is the predicate meaning True and ‘‘and’’ although a function symbol, intuitively takes two statements as arguments and returns another single conjunctive statement.[3]

This interpretation may appear to be reasonable given the English statement of the problem. But as Ohlbach discusses, this representation (along with other associated axioms) is *not* sufficient to derive the intended result.

At least part of the difficulty is not hard to see. The constant ‘‘I’’ stands for a fixed person (who is a rich knave). The point of the biconditional in **OHL**, and especially of the right-hand side, is to test whether the speaker is a rich knave, based on the ability to utter u. That is, the problem really seems to be asking, ‘‘What statement, when made by *anyone*, will convince the princess that *the person making the statement* is a rich knave?’’ The first problem with that representation is then the following: ‘‘I’’ should not be bound to a fixed individual, but should represent any ‘‘man in the street’’ who might utter u. We then suggest the alternative version:

G: $(\exists u)(\forall p)[\mathrm{CanSay}(p,u) \leftrightarrow T(\mathrm{and}(\mathrm{rich}(p),\mathrm{knave}(p)))]$

We claim to have now adequately represented the goal statement[4]; but this is still not enough. For although this goal statement expresses what we want, there are other problems arising from the truth conditions of utterances containing the pronoun ‘‘I’’. These will enter into axioms in the problem representation, rather than the goal statement.

III. Utterance Instances of Statements

This brings us to what we think is the key issue in this puzzle, an issue which has broader significance as well. Specifically, utterances are instances of statement uses, and these instances, in general, have

[2] Ohlbach’s first treatment involved the axiom $(\exists u)[\mathrm{CanSay}(I,u) \rightarrow T(\mathrm{and}(\mathrm{knave}(I),\mathrm{rich}(I)))]$ which (in addition to yielding a trivial and unhelpful answer) does not seem to correspond to his English interpretation ‘‘There exists a statement which I can say and which implies that I am really a rich knave.’’ In fact, it seems to us that the goal statement

$$(\exists u)[\mathrm{CanSay}(I,u) \,\&\, (\forall p)[\mathrm{CanSay}(p,u) \rightarrow T(\mathrm{and}(\mathrm{rich}(p),\mathrm{knave}(p)))]]$$

comes much closer to the English.

[3] Actually, a *name* of the statement.

[4] Both **G** and the second wff in footnote 2 will do equally well.

truth-values, rather than the statement in and of itself. In particular, terms in a statement may have no definite reference outside the context of an utterance. Although this concept is familiar to linguists[5] and philosophers (it is the so-called problem of indexicals which is discussed below) it is worth going into detail in the current paper, since the issue of representing knowledge in the *Knights and Knaves* problem hinges on this very phenomenon.

Typically, we think of a statement as being either true or false. This, however, is not always the case. For example, the statement:

I am a knave

will have a truth-value dependent upon who the speaker is; and so would be falsely uttered by any knight and truly by any knave.[6] Thus statements that contain indexicals (such as the word ‘‘I’’ in the above example) have meanings, and hence truth-values, that depend upon context.

Another example is the statement:

It is raining (here now)

In this case the utterance may simply be ‘‘It is raining.’’ The implication, however, is that it is raining at some particular place at some particular time. In our example, the time and place are ‘‘here’’ and ‘‘now’’ respectively, and so ‘‘here’’ and ‘‘now’’ are the indexicals that determine the truth or falsity of this statement.

Generally speaking, then, an indexical in an utterance is a sub-expression of that utterance whose meaning is determined (and thus understood) by the context in which the utterance is stated. Because of the indeterminacy of truth-values of sentences that contain indexicals, we will refer only to the truth-value of utterance-instances of such statements. An utterance-instance of a statement contains a context in which the statement was (or is) made including who the utterer is.

IV. ‘‘Who Am I?’’

[5] Including those who work in natural language processing; see for instance [Allen 1984], [Allen and Perrault 1980], [Harper and Charniak 1986].

[6] Hence, this statement can be uttered by neither knights *nor* knaves, in the *Knights and Knaves* problem!

If we look closely at any of Ohlbach's representations of the *Knights and Knaves* problem, we notice that the constant ''I'' seems to be playing two different roles. In all of his goal statements ''I'' is presumably used as the name of a particular person. For example Ohlbach's second goal statement, **OHL**, illustrates this usage. On the other hand, in the intended solution to the problem, the ''u'' of the goal statement is bound to *anriki*:[7]

and(not(rich(I)),knave(I))

where, the same symbol ''I'' appears as before but now what is of interest is its potential presence within within CanSay(I,anriki), i.e., as part of a potential utterance whose truth value depends upon who the speaker is. That is, any number of people might utter *anriki*, and its meaning would be different in each case. We now have an utterance-instance and need to know who ''I'' is before assigning a truth-value. Thus, ''I'' must be viewed as a pronoun and not a proper name here. In particular, the knighthood or ''knavehood'' of ''I'' determines the truth of *anriki*. Of course, in the world in question, only knaves (and rich ones at that) could utter *anriki*. But that is the point; the princess must be able to deduce precisely that fact: that anyone at all who utters *anriki* must consequently be a rich knave.

In what follows, we have removed this ambiguity by introducing a new predicate (TU) into the language of *Knights and Knaves*. TU is used as a 2-place predicate expression with its first argument being a person and its second an utterance. Intuitively, TU(p,u) is true if and only if u is true when uttered by person p. More precisely, we say TU(p,u) is true if and only if the substitution-instance of t resulting from replacing all occurrences of ''I'' in u by ''p'' is true. Thus the statement:

TU(John,''I am six feet tall'')

is true if and only if John (the utterer) is indeed six feet tall.[8]

V. Formalization

[7] Throughout the remainder of this paper we use ''anriki'' as a short-hand for: and(not(rich(I)),knave(I)).

[8] This is somewhat comparable to the formulation of Barwise and Perry [1983] when they speak of an utterance in a ''situation'' concerning ''I'': u[I am six feet tall]e is true (where u is the utterance ''I am six feet tall'' and e is a situation in which John is present and makes utterance u) iff John is indeed six feet tall in situation e.

We use TU as an acronym for ''truly utters''; i.e., TU(p,u) says ''p would be telling the truth if p were to utter u.''

We now introduce our notation for representing the problem. We use a first-order theory which contains the following:

I: constant (the word ‘‘I’’)

knave: function letter (knave(x) stands for the term ‘‘x is a knave’’)

rich: function letter

knight: function letter

not: function letter

and: 2-place function letter

CanSay: 2-place predicate letter (CanSay(p,u) means ‘‘p can say u’’)

TU: 2-place predicate expression (TU(p,u) means term ‘‘u’’ would be
true if occurrences of ‘‘I’’ in u are replaced by p)

T: predicate expression (T(t) means the term t is true)

Given the above notation, we can now present the axioms which will capture the *Knights and Knaves* problem as we see it. For simplicity, we suppose all variables range over knights, knaves, the princess, and utterances.[9]

All clauses we require could be derived from only three first-order axioms and one schema which are sufficient to represent the needed facts about the world in which the knights, knaves, and princess live, namely,

(1) $(\forall p)(\forall u)\{T(knave(p)) \leftrightarrow [CanSay(p,u) \leftrightarrow \neg TU(p,u)]\}$
I.e., u is a knave iff the things t that u can say are precisely those which would be false if u uttered them.

(2) $(\forall y)(\forall z)[T(and(y,z)) \leftrightarrow T(y) \& T(z)]$
This captures the meaning of the function letter ‘and’.

(3) $(\forall s)[T(s) \leftrightarrow \neg T(not(s))]$
This axiom captures the meaning of the function letter ‘not’.

[9] This follows the convention of Ohlbach. The use of either typed or relativized variables would eliminate unusual readings at the expense of more complex formulae.

(4) $(\forall p)[\mathrm{TU}(p,f(\mathrm{I})) \leftrightarrow \mathrm{T}(f(\mathrm{I}))]$

For example, this intuitively corresponds to $\mathrm{TU}(\mathrm{Bill},\mathrm{rich}(\mathrm{I})) \leftrightarrow \mathrm{Rich}(\mathrm{Bill})$, where f is ‘‘rich’’.[10]

As mentioned, axiom 4 is really a schema, and a functional one at that. So ordinary theorem provers would have to be given a mechanism to select in some fashion appropriate substitution instances. In order to avoid this added difficulty (although it should not be computationally very expensive in this case) we will continue our analysis in terms of a finite axiomatization of this schema, which requires no such mechanism.

The following four axioms recursively establish *all* possible instances of schema 4 in terms of the leftmost function symbol occurring in TU’s second argument.

$**TU** *sub and*$**:** $(\forall u)(\forall v)(\forall p)[\mathrm{TU}(p,\mathrm{and}(u,v)) \leftrightarrow \{\mathrm{TU}(p,u)\ \&\ \mathrm{TU}(p,v)\}]$

$**TU** *sub not*$**:** $(\forall u)(\forall p)[\mathrm{TU}(p,\mathrm{not}(u)) \leftrightarrow \neg\mathrm{TU}(p,u)]$

$**TU** *sub rich*$**:** $(\forall p)[\mathrm{TU}(p,\mathrm{rich}(\mathrm{I})) \leftrightarrow \mathrm{T}(\mathrm{rich}(p))]$

$**TU** *sub knave*$**:** $(\forall p)[\mathrm{TU}(p,\mathrm{knave}(\mathrm{I})) \leftrightarrow \mathrm{T}(\mathrm{knave}(p))]$

Thus the axioms we employ for the *Knights and Knaves* problem will be (1)-(3) above and the four last ones for TU, for a total of seven first-order axioms and no schemata. Below we present these axioms in clause form, leaving out those clauses that result from our axioms that are not necessary to our resolution proof.

KS1: $\neg\mathrm{T}(\mathrm{knave}(p)) \vee \neg\mathrm{CanSay}(p,u) \vee \neg\mathrm{TU}(p,u)$

KS2: $\neg\mathrm{T}(\mathrm{knave}(p)) \vee \mathrm{CanSay}(p,u) \vee \mathrm{TU}(p,u)$

KS3: $\mathrm{T}(\mathrm{knave}(p)) \vee \neg\mathrm{CanSay}(p,u) \vee \mathrm{TU}(p,u)$

A1: $\neg\mathrm{T}(\mathrm{and}(y,z)) \vee \mathrm{T}(y)$

[10] Note that this replaces Tarski’s Convention **T:** $\mathrm{T}``\alpha" \leftrightarrow \alpha$, in cases of α having the indexical ‘‘I’’.

A2: ¬T(and(y,z)) v T(z)

A3: T(and(y,z)) v ¬T(y) v ¬T(z)

N1: T(s) v T(not(s))

N2: ¬T(s) v ¬T(not(s))

TU1: TU(p,and(u,v)) v ¬TU(p,u) v ¬TU(p,v)

TU2: ¬TU(p,and(u,v)) v TU(p,u)

TU3: ¬TU(p,and(u,v)) v TU(p,v)

TU4: TU(p,not(u)) v TU (p,u)

TU5: ¬TU(p,not(u)) v ¬TU(p,u)

TU6: TU(p,rich(I)) v ¬T(rich(p))

TU7: ¬TU(p,rich(I)) v T(rich(p))

TU8: TU(p,knave(I)) v ¬T(knave(p))

TU9: ¬TU(p,knave(I))) v T(knave(p))

We are now ready for the clauses which represent our goal statement. In line with our earlier discussion, we take as our goal statement:

G: (∃u)(∀p)[CanSay(p,u) ↔ T(and(rich(p),knave(p)))]

In the clauses that follow, ‘‘g’’ is a Skolem function resulting from the elimination of the existential quantifier in the negation of **G**.

G1: CanSay(g(u),u)) v T(and(rich(g(u)), knave(g(u))))

G2: $\neg$CanSay(g(u),u) v $\neg$T(and(rich(g(u)),knave(g(u))))

Given these clauses we have been able to use resolution to give us the desired solution. Here we present our resolution proof with answer extraction showing how we come up with a solution to the *Knights and Knaves* problem. Axioms will be abbreviated using their names from section V. Clauses that are the result of a step in the proof are named (R1, R2, F13, etc.) so that we may refer to them later in the proof. For the sake of compact presentation, we abbreviate ‘and’ as ‘a’, ‘rich’ as ‘r’, ‘knave’ as ‘k’, ‘not’ as ‘n’, and ‘CanSay’ as ‘CS’. We also eliminate certain parentheses when there is no ambiguity. Key substitutions are in braces beside the resultant clause in which they appear.

MX**Resolvants** MY **Resultant Clause**
MZ R1: CS(gu,u) v T(r(gu)) v Ans(u)
& R1 → R2: CS(gu,u) v ¬T(nr(gu)) v Ans(u)
& R2 → R3: CS(gu,u) v ¬T(a(nr(gu),z)) v Ans(u)
& R3 →R4: CS(gu,u) v ¬T(nr(gu)) v ¬T(z) v Ans(u)
& R4 →R5: CS(gu,u) v ¬T(nr(gu)) v ¬TU(p,k(I)) v Ans(u)
& R5 → R6: CS(gu,u) v T(r(gu)) v ¬TU(p,k(I)) v Ans(u)
& R6 →R7: CS(gu,u) v TU(gu,r(I)) v ¬TU(p,k(I)) v Ans(u)
& R7 →R8: CS(gu,u) v ¬TU(gu,nr(I) v ¬TU(p,k(I)) v Ans(u)
& R8 →R9: CS(gu,u) v ¬TU(gu,nr(I) v ¬TU(p,a(u´,k(I))) v Ans(u)
& R9 → CS(gu,u) v ¬TU(gu,a(nr(I),v)) v ¬TU(p,a(u´,k(I))) v Ans(u)
→ CS(gu,u) v ¬TU(gu,anriki[11]) {u´,v,p → nr(I),k(I),gu}
& F11 → CS(gu,u) v CS(gu,anriki) v ¬T(k(gu)) v Ans(u)
→ CS(g(anriki),anriki) v ¬T(k(g(anriki))) v **Ans(anriki)**[12] {u → anriki}
& A2 → CS(gu,u) v T(k(gu)
& R14 → CS(g(anriki),anriki)
& R15 → T(a(r(g(anriki)),k(g(anriki)))
& R16 → T(r(g(anriki))) v ¬T(k(g(anriki))))
& R17 → T(nr(g(anriki))) v ¬T(k(g(anriki)))
& R18 →CS(g(anriki),t) v TU(g(anriki),t) v T(nr(g(anriki)))
& R19 → TU(g(anriki),anriki) v T(nr(g(anriki)))
& R20 → TU(g(anriki),nr(I)) v T(nr(g(anriki)))
& R21 → TU(g(anriki),r(I)) v T(nr(g(anriki)))
& R22 → T(r(g(anriki))) vT(nr(g(anriki)))
& R23 → T(nr(g(anriki)))
& KS1 → T(k(g(anriki))) v ¬TU(g(anriki),anriki)
& R25 → T(k(g(anriki))) v ¬TU(g(anriki),k(I)) v ¬TU(g(anriki),nr(I))
& R26 → T(k(g(anriki))) v ¬TU(g(anriki),k(I)) v TU(g(anriki),r(I))
& R27 → T(k(g(anriki))) v ¬TU(g(anriki),k(I)) v T(r(g(anriki)))
& R28 → T(k(g(anriki))) v ¬TU(g(anriki),k(I)) v ¬T(nr(g(anriki)))
& R29 → T(k(g(anriki))) v ¬TU(g(anriki),k(I))
& R30 → T(k(g(anriki)))
& KS3 → T(k(g(anriki))) v TU(g(anriki),anriki)
& R32 → TU(g(anriki),anriki)
& R33 → TU(g(anriki),k(I))
& R34→ T(k(g(anriki)))
& R35 → [ ]

VI. Discussion

Ohlbach has pointed out an interesting problem in knowledge representation. We agree in principle with his conclusion that knowledge representation is hard. In fact, if someone has to invent a new trick each time they wish to represent a problem, the task would become hopeless. Furthermore, if the language used by the AI practitioner forced the need for tricks, then there would certainly be an argument for

[11] Recall that we use *anriki* as a short-hand notation for and(not(rich(I),knave(I)).

[12] We will now drop this term from subsequent clauses, as it is not going to change and represents the desired answer.

selecting another language.

We feel, however, that neither first-order logic nor automatic theorem-proving imposes any such restriction on the *Knights and Knaves* problem. The complexity that Ohlbach discovered in trying to represent this problem is due to indexicals. In fact, his second argument of the predicate T(x,y) might be dealing with indexical-binding in some way. We have found that a proper treatment of indexical-binding makes for a natural and correct (in that a proper solution is found) representation of the *Knights and Knaves* problem.

Our solution was longer than Ohlbach's. His optimized proof had 20 steps, while ours has 36. Thus the new issues we have introduced into the problem representation have not reduced the complexity; rather they have increased it, but not excessively so. Thus the use of indexicals seems viable within an automatic theorem-proving context.

Furthermore, we feel that this problem is indicative of a whole class of problems that can be handled in a similar fashion, i.e., not dependent upon isolated or ad hoc tricks. In the *Knights and Knaves* problem we defined TU in terms of the indexical ''I'' only. This is because ''I'' is the only indexical of importance in this problem. In broader contexts, however, this would be insufficient and generalizations of TU would be necessary. Thus, we offer TU as a step toward a uniform solution to the problem of automatic theorem-proving with indexicals. It will be interesting to see how well generalizations of TU handle other indexicals and other problems.

Acknowledgments

This research has been supported in part by the following institutions:

National Science Foundation Grant #OIR-8500108
The U.S. Army Research Office (DAAG29-85-K-0177)
The Martin Marietta Corporation

We would also like to thank Harold Edmundson, Amy Weinberg, Mike Mazurek, Alan Frisch, Terry Gaasterland, Dominic Duggan, Rosalie Hall, Andrea Marra, Jack Minker, Jim Hendler, and Annie Gal for their helpful observations.

Bibliography

(1) Allen, J. [1984] Toward a general theory of action and time, *Artificial Intelligence*, Vol. 23, 123-154.

(2) Allen, J. and Perrault, R. [1980] Analyzing intentions in utterances, *Artificial Intelligence*, Vol. 15, 143-178.

(3) Barwise, J. and Perry, J. [1983] *Situations and Attitudes*. MIT Press, Cambridge, Massachusetts.

(4) Harper, M. and Charniak, E. [1986] Time and tense in English, *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, Columbia University, New York, 3-9.

(5) Ohlbach, H. J. [1985] Predicate logic hacker tricks, *J. of Automated Reasoning* Vol. 1, No. 4, 435-440.

(6) Perlis, D. [1985] Languages with self reference I, *Artificial Intelligence*, Vol. 25, 301-322.

(7) Smith, B. [1986] Varieties of self-reference. *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning About Knowledge,* ed. J. Halpern, Morgan Kaufmann, Los Altos, California.

(8) Smullyan, R. [1978] *What is the Name of this Book?*. Prentice-Hall, Englewood Cliffs, New Jersey.