DIALOGUES BETWEEN HUMANS AND AN AETIFICIAL BELIEF SYSTEM

Kenneth Mark Colby and David Canfield Smith

Computer Science Department
Stanford University
Stanford, California

An artificial belief system capable of conducting on-line dialogues with humans has been constructed. It accepts information, answers questions and establishes a credibility for the information it acquires and for its human informants. Beginning with beliefs of high credibility from a highly believed source, the system is being subjected to the experience of dialogues with other humans.

Introduction

Imagine an extraterrestrial entity sent to earth to collect information from human beings. It has no senses, as we know them, and it cannot move about. Its only means of communication is through written language. Thus the information it acquires will be dependent upon what it is told by human beings in written dialogues. After it has participated in dialogues with a number of humans, what would it believe?

We have realized this imagined situation by creating an artificial belief system (ABS}) in the form of a computer program. Our goal was to study certain properties of credibility functions in a synthesized artificial system whose structure and starting conditions were entirely under our control. Before describing the artifact and its behavior, perhaps something should be sa i.d abuut the research strategy involved and its relevance to artificial intelligence.

One synthesizes an imaginary system to .". simplity a computer problem and to bring variability under control. In the case of credibility processes in humans we are dealing with complex intelligent behavior difficult to control and manipulate. Hence, one would like to work with a simpler system showing some of the essential properties of a human belief system.

Given that one creates such a system to achieve exactness, simplicity and conciseness, how is the artifact to be manipulated or set in conceptual motion? Artificial intelligence provides a way of reversing the traditional technique of experimentation with humans. Instead of placing a person under the artificial conditions of a laboratory experiment, we can subject an artifjeial system to the naturally-occurring conditions of human linguistic communication. Our purpose is to study the processes of credibility functions in a synthesized artifact designed to perform the intelligent task of arriving at states of belief or disbelief about what it is told by persons in dialogues.

Some parallels between these processes in an artificial belief system and in human credibility development will be apparent in later sections. However we are not here attempting to simulate a human belief system. Such el torts *are* described elsewhere. 1,2

We shall begin with a general description of the artifact ABS}• The subscript 'one[1] indicates this is the first version of a class of anticipated models.

Starting Conditions

ABS_1 is a computer program, written in MLISP, which runs in an interactive mode on the PDP-6/IO time-sharing system of the Stanford Artificial Intelligence Project. MLISP is a list processing language which translates Algol-like M-cxpressions into the S-expressions of Lisp 1..5.3.4

The program instructions of ABS1 consist of rules of handling input strings and for generating output replies. When a human informant communicates with ABS_1 from a terminal, he types whatever he wishes using any vocabulary names he chooses. In order to avoid many of the complexities of natural language expressions, the syntax is restricted to certain forms.

The first form is termed a statement. Its pattern is Subject-field, Verb-field, Modifier-field. The subject and modifier fields can contain any number of words of any type except linking verbs. The verb field is limited to third person singular forms in three tenses of linking verbs (be, seem, appear, feel, become). The verb can be optionally preceded by an auxiliary (can, could, must, ought, would, should, shall, will) and followed by a modal operator (certainly, probably, possibly) plus a determiner (a, an, the)• The negation 'not' is also permitted. Modal operators allow persons to express degrees and directional strength of relations about a probablistic world. A statement ends with a period. Here are some examples of statements:

(1) John is a man.

(2) John is certainly not an intellectual.

(3) My brother John was a truck driver.

(4) Young Peter will probably not be e problem in school.

(5) Disease seems to be a cause of death.

Two interrogative forms are permitted: Subject-Verb-Modifier or Verb-Subject-Modifier, followed by a question mark. These are examples of questions.

(1) Is John an intellectual?

(2) John is not an intellectual?

A fourth syntactical form allows expectancy or implication rules in the format x (verb-field) (modifier-field) implies x (verb-field) (modifier-field) followed by a period. For example:

(1) x is a man implies x is certainly a person.

(2) x is an Italian implies x is probably a Catholic.

(3) x is a white Southerner implies x is possibly a racist.

(4) x is a $Bill4 implies x is a WASP.

The term 'implies' here does not refer to logical implication. These rules correspond to the expectancies of psychological implication in which, given that one situation is the case, a human expects a second situation to be the case.

Rules contain two sorts of variables. The variable 'x' stands for 'anything'. A variable beginning with a dollar-sign followed by the informants name and a number is termed a constrained variable. It is defined by an informant as he chooses. Thus the conjunctive definition of $Bill4 might be:

$Bill4 is a white man.

$Bill4 is an Anglo Saxon.

$Bill4j- is a Protestant.

For a person to be substituted into the variable $Bill4, facts about him would have to fit the requirements stated. Constrained variables can be nested. Disjunctive definitions are handled by a function ANY( ) whose argument determines the number of facts required to meet the requirements of the constrained variable. Constrained variables allow implicational rules to be of arbitrary complexity while at the top level they remain a single expression consisting of two expressions linked by the term 'implies'.

When input in any of these four forms is typed in, the program undertakes a number of operations. At the very start ABS1 consists only of program rules with no data. All eventual data derive from subsequent written communication with humans. When ABS_1 begins its dialogue with a person, it requires his name, and recognizes it if he has been conversed with previously. If so, it simply prints out 'Go ahead'. If not, it prints out instructions regarding its allowable formats. We shall describe what happens to each type of input during the conversational phase termed 'Talktime'.

Statements and Rules

A statement is characterized by its format and by a period as its terminal symbol. Upon entry a statement is put on the statement list of the current informant engaging in the dialogue. A reply is then output to the informant which indicates that the input has been received and that ABS1 awaits the next input. Thus in handling statements during Talktime, ABS_1 simply absorbs them in a sponge-like fashion, stores them in its data base and indicates that it is ready for more information. It prehends what it is told in the form of statements and considers them to be facts of observation.

A rule is characterized by its format in which two expressions containing variables are connected by the term 'implies'. On entry a rule is put on the rule list of the informant. The reply 'Rule O.K.' is output and ABS_1 awaits further input.

Questions

When an informant enters an expression in the form of a question, ABS_1 stores it on the informant's Question list and then seeks to answer the question. Suppose the question was:

Q(l) 'IS Wallace a racist?'

ABS_1 first searches the questioner's statement list of expressions looking for two sorts of statements, (a) a direct identity or negation statement or (b) a similar statement which could answer the question. A direct identity statement would be:

S(l) 'Wallace is a racist.'

A direct negation statement would be:

S(2) 'Wallace is not a racist.'

Similar statements, in which the verb-field content of the input question and the statement found are considered similar, would be:

S(3) 'Wallace was probably a racist.'

S(4) 'Wallace seems to be a racist.'

S(5) 'Wallace is possibly a racist.'

or any of these statements containing the negation 'not[1].

If S(l) is found on the informant's statement list ABS_1 outputs the reply:

R(l) 'Yes, Wallace is a racist (according to --).'

where the blank is filled in by the informant's name. If S(2) is found the reply is:

R(2) 'NO, Wallace is not a racist (according to --).'

If both statements, i.e., a contradiction, existed on the informant's statement list, the first one found is used for a reply. The statement list is organized with the most recent expressions at the top and ABS_1 considers the most recent of two contradictory statements to be the more representative of the informant's current state. Such a principle also holds for similar statements. That is, if a similar statement such as S(4) is found first the reply would be:

R(3) 'NO. but Wallace seems to be a racist (according to --).'

even though a direct identity or negation might exist deeper down in the statement list. Thus ABS_1 allows more recent statements to count more heavily in answering questions.

Another type of expression on the statement list can qualify as a reply to Q(1). This is a statement such as

S(6) 'Wallace is probably a racist. (Previous conclusion)'.

which was not input by the informant but which ABSj concluded to be a fact from an earlier conversation with this particular informant. How such a conclusion is generated will become clear by considering what happens when statements such as S(l)-S(6) are not found.

The search up to this point has been concerned with statements, i.e., facts. If direct or similar facts are lacking, $_{ABS1}$ turns to the use of rules and attempts to infer a factual conclusion which could answer the question. In replying to Q(l) the relevant implication rules would be ones containing the term 'racist' such as:

l(l) x is a WASP Southerner implies x is probably a racist

or

1(2) x is a racist implies x is certainly a bigot.

If l(l) in which the term 'racist' appears on the right-side of the rule, is found on the informant's rule list, a search is made among the facts to see if the left-side of the rule can be satisfied with 'Wallace' being the value of the subject variable x . If the fact

S(7) Wallace is a WASP Southerner

is found, then the fact 'Wallace is a racist' is generated and stored on the statement list of the informant as a conclusion made by ABS_1 while conversing with the informant. The reply to Q(l) would then be:

R(4) 'Yes, Wallace is probably a racist because x is a Wasp Southerner implies x is probably a racist (according to --) and Wallace is a Wasp Southerner (according to --).'

Using an implicational rule from left to right is in the direction of finding strong confirmatory evidence. Another inference process moves from right to left in a direction of what is considered to be weaker supportive evidence. If the rule l(2) were found along with the fact:

S(8) 'Wallace is certainly a bigot.'

then a weaker conclusion would be generated and the reply would be

R(5) 'Wallace is possibly a racist because x is a racist implies x is certainly a bigot (according to --) and Wallace is certainly a bigot (according to --).'

A number of rules can be combined through backward chaining to establish a conclusion. The number of rules used depends on a parameter which can be set to yield an inference process of any desired depth. For example, suppose the rules were:

1(3) x is a red-neck implies x is a lower middle class Southerner.

l(i|-) x is a lower middle class Southerner implies x is a hater of negroes.

1(5) x is a hater of Negroes implies x is a racist

If the input question was:

Q(2) 'Is Maddox a racist'?

the rule l(5) would be found and then an attempt made to find the fact 'Maddox is a hater of Negroes'. If this search fails, ABS_1 backs up to rule *l(4)* and a search made for a factual match of the left side of the rule. If that fails, ABS_1 backs up to rule 1(3). If the fact 'Maddox is a red neck' is found then the conclusion that 'Maddox is a racist' would be established. The conclusion is output as a reply to Q(2) along with the chain of reasoning and facts used to reach this conclusion.

It is important to notice that while the inference processes of ABS_1 is 'logical' in the sense of using valid reasoning with various degrees fo strength, the semantic or conceptual nature of the expectancy rules may determine a conclusion

which mig^it seem peculiar to the pure logician. For instance if the rule acquired from an informant stated:

I(6) x is a racist implies x is a Wasp Southerner

and if the fact were found that 'Smith is a racist' then the conclusion 'Smith is a Wasp Southerner' would he reached. However, if it is Ian Smith of Rhodesia who is a racist, the conclusion would be in error,

ABS_1 has no control over what people tell it. If it acquires expectancy rules like 1(6) then it uses them in its reasoning. While its inference process is formally valid, it can come to empirically incorrect conclusions because of the conceptual content of its facts and rules. Such a situation has obvious implications for human reasoning.

Thus far we have described the processes ABS_1 undertakes in replying to a question from an informant using only that informant's facts and rules. After ABS-. has conversed with several persons, it contains lists for each of them in its data base. When asked a question, ABS_1 first searches the lists of the current informant. If a reply cannot be found or generated, a search is made of the lists of each informant. Facts from one informant can be combined with rules of another to yield a conclusion. If informants have input contradictory expressions, both are used in the reply. An input fact is considered stronger than one inferred from rules and, as mentioned, a more recently received fact is considered more representative than an earlier fact from an informant.

Questiontime

ABS1 During the conversational phase of Talktime, intakes statements and rules and replies to questions. When an informant has finished what he wants to say, he types the word 'Done' which initiates a phase in which he in turn is asked some questions.

ABS-, searches the statement list of the informant looking for statements in which the Subject-fields 8re identical but the Modifier-fields differ. When such a pair is found, e.g.:

S(9) 'Sam is a baseball player.'

S(10) 'Sam is a man.'

the question is formed:

•Does being a baseball player imply being a man?'

A check is made to see if the question has been asked before or whether the rule it refers to is already present. If not, the question is output to the informant who can reply with 'yes', 'certainly*, 'probably', 'possibly' or 'no'. If an affirmative reply is received, ABS[_ stores the rule:

'x is a baseball player implies x is a man' with any received modal operator.

If the reply from the informant is 'no', a second question involving a negation is asked. In this case it would be: D*oes* being a baseball player imply not being a man?' If an affirmative is received, the negation containing rule is saved. If the reply is 'no', the rule is discarded. The asymmetry between affirmation and denial stems from the obvious but little-honored ambiguity of 'no' in human communication. When a human denies some expression it is difficult to know whether he is referring to a negation or the complement of a set.

Since the data base lists are temporally ordered S(IO) might be found before S(9)• Hence the converse questions would have been asked first. Both questions ore asked in order to pick up any possible relations between the two modifier fields.

The number of questions asked an informant-can be quite large and humans tend to find answering tedious. Hence one can stop the rule questions at any time by typing 'Quit'. On receiving tne latter symbol, ABS_1 attempts to engage the informant in a process of categorizing his concepts. There are six domains of interest or semantic categories ABS_1 is interested in - Politics, Religion, Student Dissent, Race, Persons and Other, the latter meaning not belonging to one of the first five. Subject and modifier fields arc searched for concepts (excluding variables, determiners, modals and prepositions) not already categorized in a domain. When one is found it is typed out and the informant is asked to assign the concept to one or more domains with which he judges it most concerned. Also the informant is given an opportunity to recategorize a concept he previously categorized. As with questions the informant can bring an end to this mode of interaction by typing 'Quit'. ABS_1 then enters a phase of processing called 'Thinktime' in which it attempts to conclude new facts from what it has been told and to establish a credibility for not only the facts and rules it has received but also for the informant as a believable source of information.

Thinktime

In the section on Questions we described how ABS1 can form a conclusion in response to a question from an informant. Combining facts and rules from one or more informants, it undertakes chains of inference in order to reach a conclusion which satisfies the input question. During Thinktime a similar process is carried out using all possible facts and rules. This is done after each conversation with each informant since new information can permit new conclusions to be reached.

Hoping for clarity of exposition we have thus far described the operations of ABS_1 without referring

to its chief characteristic, namely, that this is a belief system. A belief system, whether naturally occurring (as in humans) or artificial, assigns credibility to the propositions it holds and to sources of information. Our main purpose in constructing this artifact was to study these credibility processes in an artificial system which interacted with humans.

In ABS_1 credibility is a weighted additive function of foundation and consistency. The foundation of a proposition (p) is the weighted average of the credibilities of beliefs (Bi) in the system which imply the proposition (p) or its negation.

$$\text{FOUNDATION (p)} = \frac{\sum(Bi \Rightarrow p)}{\sum(Bi \Rightarrow p) + \sum(Bi \Rightarrow \neg p)} \quad \text{Eq(1)}$$

The number of beliefs considered is determined by a depth parameter which determines the amount of evidence in arriving at the foundation for a belief. Consistency is the converse of foundation:

$$\text{CONSISTENCY (p)} = \frac{\sum(p \Rightarrow Bi)}{\sum(p \Rightarrow Bi) + \sum(p \Rightarrow \neg Bi)} \quad \text{Eq(2)}$$

The formula for credibility is:

$$\text{CREDIBILITY} = \text{PCRED} + \text{ALPHA} \times (W \times (\text{FOUND} - 0.5) + (1-W) \times (\text{CONSIS} - 0.5)) \quad \text{Eq(3)}$$

where ALPHA is a weight relating the degree of importance between the credibility of a source and a particular belief, W is a weight representing a degree of importance which foundation and consistency have in relation to one another. Initially ALPHA was set to the value 0.2 and W to 0.8. Thus, for *AB£>i,* the foundation for a belief is four times more important than the consistency. Using various values for these parameters we can carry out ideal experiments in which only one variable at a time is manipulated and exact replication can be achieved.

In the formula for credibility PCRED stands for a preliminary credibility. It is clear that for the artifact to get off the gound it must have some initial credibility assignments. To start it off we allowed ABS_1 to assign a high credibility to the informant KMC and to everything he told it without examining this body of data for its degree of foundation or consistency. This situation is analogous to that of a human child who early in life believes everything a parent tells him. For sociogenetic transmission of information to occur it is necessary that a child be a credulous system, willing to accept as true that which he recieves as testimony from authorities such as parents. In time a child may come to question these authorities when he receives contradictory information from peers, other authorities and from his own personal experience with the world.

To study the process of change in ABS^1 we began with KMC as a highly credible source and with his input in all domains of interest being accepted as strongly credible. Each new informant is initially given a global credibility and a domain credibility of 0.5« The bipolar scale of credibility used is as follows:

CREDIBLE	STRONG	0.9
	MEDIUM	0.7
	WEAK	0.6
NEUTRAL		0.41 - 0.59
INCREDIBLE	WEAK	0.40
	MEDIUM	0.30
	STRONG	0.10

Since these numbers really represent rank orders they are rounded off at the end of numerical computations to yield the nearest value. It is to be noted that there is a greater distance between strong-medium than between medium-weak. A credibility value in the neutral range means the proposition is held temporarily as a candidate for belief. After ABS_1 proceeds through the phase called Thinktime, there will exist no credibilities in the neutral range. ABS_1 strives to polarize its beliefs and thus tends to avoid states of neutrality or doubt.

During Thinktime, ABS1 attempts to establish the credibility of the statements of informant with whom it has most recently conversed. It first assigns a preliminary credibility of each of the new statements based on the general credibility of this informant in the semantic categories or domains to which the statement belongs. This becomes the PCRED of Eq(3). It then looks at all relevant statements and rules which already have credibilities to compute the foundation as in Eq(l) and consistency as in Eq(2). Finally it assigns a credibility to a statement according to Eq(5).

As has been described, during Talktime ABS_1 uses an inference process of combining statements and rules to reply to questions whose answers cannot be found as already present statements. During Thinktime, ABS1 uses this inference process to generate new statements. All possible conclusions are drawn using both a left to right and right to left inference process on the rules. The statements concluded are assigned to the atom 'self' which is a member of the Persons list. Thus 'self' is treated as an informant like any other and 'self' can converse with ABS1. The statements generated by 'self' in the inference process of Thinktime are assigned credibilities and 'self in turn receives global as well as domain credibilities. This type of selfhood has amusing consequences, such as finding 'self' to be incredible in some domains or judging the statement 'self is not a machine' to be credible.

The final procedures of this phase involve a reevaluation of the credibilities of the most recent informant as a source and of 'self'. Since informants tend to repeat themselves over time, the frequency of a statement or a rule is here used

to compute the credibility of an informant in each of the domains of interest which in turn allows a new global credibility to be assigned to the informant. The same process is carried out on 'self. ABS1 does not then try to recompute credibilities of statements and rules from other informants. The potential circularities here are evident and must be avoided.

Discussion

We have described in detail an artificial belief system (ABS1) in the form of a computer program which attempts to perform the intelligent task of estimating the credibility of human sources of information. There are many uses for such a program. At the moment we are using it to study the problem of change in a belief system. Given starting beliefs of high credibility from a highly credible initial source, ABS1 is being exposed to dialogues with other human sources who may agree or disagree with the initial source. Does ABS1 change as a result of this experience and what is the change a function of? We shall answer these questions in a future report.

Summary

An artificial belief system (ABS_1) capable of conducting dialogues with humans has been constructed. It intakes whatever information it receives, answers questions and establishes the credibility of the information as well as its human source. It is currently being used to study the problem of change and resistance to change in a belief system.

References

[1] Colby, K. M. Computer Simulation of Change in Personal Belief Systems. Behavioral Science, 12, 248 253 (1967).

[2] Colby, K. M., Tesler, L. and Enea, H. Experiments with the Data Base of a Human Belief Structure. (See this volume.)

[3] Enea, H. MLISP. Stanford Computer Science Department Technical Report, No. 92, March 14, 1968.

[4] Smith, D. C. MLISP Users Manual. Stanford Artificial Intelligence Memo No. 1969. (In preparation.)

Acknowledgements

This research is supported by Grant PHS MH 066-45-07 from the National Institute of Mental Health, by (in part) Research Scientist Award (No. K-14,433) from the National Institute of Mental Health to the senior author and (in part) by the Advanced Research Projects Agency of the Office of the Secretary of Defense (SD-I83).