Building A Priori Knowledge Spaces from Interrogative Domains

Cameron A. Hughes  
Ctest Laboratories  
Youngstown, Ohio, USA  
cameronhughes@ctestlabs.org

Tracey D. Hughes  
Ctest Laboratories  
Youngstown, Ohio, USA  
traceyhughes@ctestlabs.org

Abstract—A Priori knowledge is the knowledge an agent has gained prior to experiences and learning. A priori knowledge acquisition along with relevant functional ontology building remain obstacles in the process of building knowledge-based agent-oriented systems. In this paper, we describe epistemic analysis techniques that we are exploring at Ctest Laboratories that are used to automatically discover ontological artifacts within the transcripts of interrogative domains. These ontological artifacts are then used as the fundamental basis of a priori knowledge spaces for epistemic agent-oriented systems. We are using ROGUE (Real-time Ontology Generation Using Epistemic Agents) to perform the transcript and text mining. ROGUE is a multi-agent system under development at Ctest Labs.

Keywords; epistemology, agents, epistemic structures, transcript mining, text mining, propositional knowledge, interrogative transcript, epistemological analysis

I. Introduction

In this paper, we will present some of our work at Ctest Laboratories where we are using interrogative domains as knowledge sources for building the a priori knowledge space of epistemic agents. We are interested in interrogative domains of digital transcripts because of the sheer scope of the subject matter. The semi-structured nature of interrogative domains can assist in resolving some of the ambiguities involved in natural language processing. Quality and timely knowledge acquisition along with relevant functional ontology building continue to be bottlenecks to the process of building knowledge-based agent-oriented systems. While there are many efforts underway to remedy this [1], knowledge acquisition and ontology building remain obstacles [2]. Our notion is we can dispense with traditional knowledge acquisition techniques and use transcript mining to build a priori knowledge spaces of the epistemic agents.

II. Tripartite Analysis and A Priori Knowledge

We use the Tripartite Analysis [5] of knowledge as the basis for our epistemic agents. Although the Tripartite Analysis has shortcomings and has been thoroughly criticized, see [6] for the basic attack on the Tripartite deconstruction, it is well suited for our implementation of epistemic agents. In this analysis, propositional knowledge is understood as justified true belief. This is in contrast to the popular use of knowledge as information in context. We intersect the Tripartite Analysis of knowledge and our epistemic agent by using Kripke structures [7]. If we let \( M = (S, \pi, K_1,...,K_n) \) then \( K \), what the agent knows, and \( S \), the worlds the agent considers possible, are used to represent the notion of modal truth in the context of the Tripartite Analysis. Further, \( S \) is expanded and defined by our agent's epistemic structure [8].

We take our cue for our concepts of a priori and posteriori knowledge from [5] and [6]. In short a priori knowledge that comes before experience and learning (the initial state of knowledge for the agent). We have adjusted the notions so that they apply to computer agents. So that a priori knowledge is the original knowledge that the agent starts with prior to execution for the first time. This is in contrast to the knowledge the agent learns from experience, induction, or inference [16].

III. Interrogative Domains as Knowledge Sources

In any type of knowledge acquisition from text there is a concern for truthfulness and the validation of truth. The interrogative domains that we are concerned with are digital transcripts of courtroom trials, congressional hearings, and law enforcement interrogations. We chose this literary form as a source of propositional knowledge because in each there is either an implicit or explicit search for the truth and validation of truth. In some cases the goal of these interrogative domains is to substantiate facts, information, or knowledge that is available and in other cases the goal is to discover information or knowledge that is not readily available.

In natural language processing or text mining context, it is not always clear when or how one grammatical item is related to another, when the scope or focus has changed, or when new referents have been introduced and old ones dropped. This leads to ambiguity and is part of what makes natural language processing challenging [13]. Each type of domain utilizes mechanisms in order to handle ambiguity considered an obstacle to understanding. In courtroom transcripts, the relationship between a question and answer must be predetermined and clear. Questions or answers that are ambiguous or vague will be objected to or asked to be clarified by the presiding judge. The Q&A pairs in the transcript greatly simplify the segmenting task [14].

What is also interesting about these transcripts is the sheer scope of the subject matter. Transcripts of civil proceedings can range from expert testimony on the effects of carcinogenic chemicals (known as PCB) in fish
oil dietary supplements to the structure of the human genome. It is due to that range of domains that we believe useful ontological artifacts can be extracted.

IV. Epistemic Agent Definition

Epistemic agents are distinguished from other types of agents by the fact that they have an epistemic structure [8] and their agent cycle is focused on knowledge acquisition, justification, knowledge maintenance, and proposition classification. While there are several classic notions of what constitutes an agent cycle, see Shoham [9], Rao [10] and Wooldridge [11], we take our cue from Kowalski's and Sadri's work in logic programming and multi-agent systems [12]. Further, epistemic agents have a DNA (Deductive Nuclear Architecture) [4] as opposed to a BDI (Belief Desire Intentions) architecture [11]. The DNA is a structure designated by

\[ \alpha = < E_s, \Theta, I, \delta > \]  

where:

- \( E_s \) is the Epistemic Structure,
- \( \Theta \) is deductive theory of inference over \( E_s \) and \( I \),
- \( I \) is a set of interrogative types,
- \( \delta \) is the deductive response (the set of conclusions reachable from \( E_s \)).

The epistemic structure \( E_s \) is a knowledge structure in which its purpose is to hold or store the knowledge (justified true belief) of an epistemic agent [8] defined as:

\[ E_s = < G_1, G_2, J, V, F > \]  

where:

- \( G_1 \) is a Graph of a priori propositions,
- \( G_2 \) is a Graph of a posteriori propositions,
- \( J \) is a set of justification propositions,
- \( V \) is a vector of Commitment,
- \( F \) is a non monotonic truth maintenance function on \( E_s \).

V. A Priori Knowledge Space of the Epistemic Agent

The a priori knowledge space consists of the domains of propositional knowledge extracted from the transcripts. If we let \( d = \{ E_s1, E_s2, E_s3, ... E_sn \} \), where \( d \) is a set of epistemic structures representing a particular domain or a collection of domains. Then we have:

\[ K_s = \bigcup_{i=1}^{N} d \]  

where \( K_s \) is epistemic reality or the total knowledge space of the agent and \( K_s \) is a set union of the domains. Typically \( K_s \) is built using one or more of the standard knowledge acquisition techniques such as conducting interviews, disseminating questionnaires, repository evaluation, and local-remote observation. Rather than acquiring the base ontology through the aforementioned traditional processes [3], we are exploring the use of transcript mining [4] to automatically identify the base ontology for the agent’s a priori knowledge space.

Once \( K_s \) is built, we can summarize the behavior of the epistemic agents \( e_n \) if we let:

\[ e_n / \alpha \]  

be a set of agents with DNA then the Fundamental Epistemic Axiom can be stated as:

\[ \forall q \delta(q) \rightarrow K_s \vdash \delta(q) \subseteq e_n / \alpha \Theta(I, (q)) \cap M \]  

This axiom is used to guarantee the epistemological soundness of the work performed by the epistemic agents. Fig. 1 shows the block diagram of one of the ROGUE agent architectures that we use at Ctest Laboratories. The agent exists in an environment from which it senses percepts (inputs), performs actions, and then outputs to the environment. For example, the ILP Q&A forms is an inductive logic program that processes unknown Q&A pairs. It updates the semantic and propositional knowledge component entailed and inferred propositions and updates the GRAMMAR (ISIS-Interrogative Sentence Interface Subsystem) with new grammatical forms. ISIS provides a question-based NLI. The ROGUE

![Figure 1. A block diagram of the architecture of the ROGUE agent.](image-url)
agent outputs to the Unknown Words component and to the COGNOPAEDIA which is an ontology and knowledge construct that replaces the application domain.

VI. PROPOSITIONAL KNOWLEDGE FROM INTERROGATIVE ENTAILMENT

We can take advantage of interrogative entailment for transcript mining purposes. Most question and answer pairs entail one or more statements. For example, Table I shows a question and answer pair taken from one of the courtroom trial transcripts that we used in our datasets. The simple semantic meaning of the Q1 combined with A1 semantically entails statements S1a and from S1a we may infer S1b and S1c.

VII. ONTOLOGICAL ARTIFACTS

The ontological artifacts we are interested in developing are the vocabulary, base relationships, and the primary entities and objects of the ontology taken directly from the propositional knowledge. That which exists or everything that is under discussion, i.e. the ontology [15], can be found in the transcript. When it comes to truth analysis, the epistemic reality(4) and the worlds the agent can be found in the transcript. When it comes to truth everything that is under discussion, i.e. the ontology [15], from the propositional knowledge. That which exists or primary entities and objects of the ontology taken directly developing are the 

Step 1: Segment the transcript into blocks of question and answer pairs while maintaining the chronological appearance of witnesses, attorney's evidence, etc.

Step 2: Classify the question and answer pairs according to the 13 basic categories of questions and answers.

Step 3: Resolve anaphora and mark substitutions between the question and answer pair blocks.

Step 4: Using entailment, convert the question and answer pairs to propositions.

Step 5: Using the predicate simple_subject(), simple Predicate(), simple_object() to extract the ontological artifacts from the propositions.

Step 2's classification has important ramifications because if the question and answer pair is classified as a location type, then the entailed proposition will also be a location type. That is the proposition will make an assertion about location. It is also important to do the Q&A classification at this point because it helps with the anaphora resolution rules.

If we let \( T = \{\text{Set of Q&A in transcript}\} \) then:

\[
T \vdash T_1
\]  
(7)

TABLE I. ENTAILED AND INFERRED PROPOSITIONS FROM A Q&A PAIR.

<table>
<thead>
<tr>
<th>Question &amp; Answer</th>
<th>Entailed Proposition</th>
<th>Inferred Proposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. And you saw her take us all to a site that was in London, is that correct?</td>
<td>S1a. I saw her take you to a site that was in London.</td>
<td>S1b. She took us all to a site in London.</td>
</tr>
<tr>
<td>A1. Yes.</td>
<td>S1c. We all went to a site in London.</td>
<td></td>
</tr>
</tbody>
</table>

where \( T_1 \) is the set of propositions entailed from \( T \). Once we have \( T_1 \), we perform a model theoretic analysis on \( T_1 \) to extract the ontological artifacts that we are interested in. This analysis is done using three predicates from the ROGUE system. The predicates are:

\[ \text{simple_subject}(P_i), \text{simple_predicate}(P_i), \text{simple_object}(P_i) \]

where \( P_i \) is an interrogatively entailed proposition. And the three simple predicates take \( P_i \) as arguments and produce the corresponding ontological artifact. The ontological artifacts are captured by the model theoretic semantics of the transcript. Here the model is described as \( m = (D,f) \) where \( m \) is a model theoretic semantic representation of all the language that is contained in the trial corpus. \( m \) consists of a pair \( (D,f) \) where \( D \) is the Domain which is the set of people and things referenced in the corpus (e.g. defendants, jurors, attorneys, witnesses) plus the relations (e.g. lawyer(X,Y), trial_day(N), etc.) between those people and things. \( f \) is an interpretation function which maps everything in the language onto something in the domain [14]. \( f \) is implemented using the Lambda Calculus operator \( \lambda \) and \( \beta \)-conversion. \( m \) captures completely our ontological artifacts. It is important to note here that the process that extracts \( m \) from \( T_1 \) requires robust natural language processing (NLP). At Ctest Laboratories, we use ISIS which provides the question-based natural language interface.

VIII. A PRIORI KNOWLEDGE SPACE

Recall \( S \) from (1) which are the worlds that the agent considers possible. In the context of our Kripke structures for the agent we let \( S \approx T_1 \) which gives us a substitution for our Kripke structure:

\[
M = (T_1, \pi, K_1, ..., K_n)
\]  
(9)

This substitution allows us to relate what the agent knows and what the agent considers possible to the epistemic reality or closed world of the digital transcript. These a priori propositions are stored in \( G_1 \) of \( E \) and constitute the a priori knowledge space of the agent. The ROGUE system automatically generates \( T_1 \) using interrogative entailment [4]. The propositions in \( G_1 \) are posteriori because they are learned during the agent cycle.
and by induction. Initially $G_2$ is $\{\}$, and we have $G_i$ with a cardinality $>0$. So initially prior experience and learning $K_i$ from (4) is the a priori knowledge space of the epistemic agent.

Further, from (4) we know that $K_i$ consists of 1 to $n$ domains $(d)$ where each domain consists of 1 to $n$ $E_d$. Then the a priori $K_i$ will have $nE_i$ that are partially constructed. They are partially constructed because the ontological artifacts that are mined from the transcript will only populate $G_i$ and in some cases some of $J_s$ and parts of $V_c$. Some of the a priori propositions will serve as justifications and as a basis for the agent’s level of commitment [4] to the other propositions in $G_i$ and to those that will later populate $G_2$. It is $E_i$ namely $J_s$, $V_c$, and $F$ which distinguishes the epistemic agent’s knowledge space from the traditional knowledgebase.

So the a priori knowledge space is built strictly from entailed propositions taken directly from the Q&A pairs of the digital transcript using only entailment, deduction and logical implication. And therefore the responses the epistemic agent is capable of giving relies on logical inference and we are guaranteed:

$$\forall q \delta(q)$$

from (6) namely we can rely on the agent’s answers given that the original propositions are not possibly false in all the worlds the agent considers possible.

IX. Future Work

So far we’ve have only built knowledge spaces for a single domain extracted from our dataset of digital transcripts. We are in the process of attempting multi-domain knowledge spaces. Besides the problems of NLP and perform issues, there is an access issue for courtroom digital transcripts. Many infamous transcripts are readily available in digital form but it has been difficult to access large numbers of transcripts. There is no consistency of availability across state districts. Some districts will make available transcripts at a high costs but they are not in digital form. While the digital transcript as a knowledge source on the surface appears to be very fertile ground, we still have much work to do. We believe that access to larger numbers of transcripts will help the transcript mining process produce multi-domain models. Currently, we are using datasets of less than 100 transcripts. It appears that the more transcripts we consider on a topic the higher the knowledge credential and payoff. We are already looking at ways that we can enhance our ISIS NLP system as a result of its failure to process 100% of question and answer pairs presented to it. The failure rate was about 30% for courtroom transcripts but we had failure rates of up to 57% for congressional hearings. We still see a great deal of promise in using digital transcripts to automate parts or all of the knowledge space building process.