The Comprehensive Terrorism Knowledge Base in Cyc

Chris Deaton, Blake Shepard, Charles Klein,
Corrinne Mayans, Brett Summers, Antoine Brusseau, Michael Witbrock, Doug Lenat
Cycorp, Inc., 3721 Executive Center Drive, Suite 100, Austin, Texas, 78731, USA
{cdeaton,blake,klein,cmayans,bsummers,brusseau,witbrock,lenat}@cyc.com

Keywords: Question Answering, Structured Argumentation, Social Analysis, Temporal Analysis, terrorism

Abstract

This paper describes the comprehensive Terrorism Knowledge Base™ (TKB™) which will ultimately contain all relevant knowledge about terrorist groups, their members, leaders, affiliations, etc., and full descriptions of specific terrorist events. Led by world-class experts in terrorism, knowledge enterers have, with simple tools, been building the TKB at the rate of up to 100 assertions per person-hour. The knowledge is stored in a manner suitable for computer understanding and reasoning. The TKB also utilizes its reasoning modules to integrate data and correlate observations, generate scenarios, answer questions and compose explanations.

1. Introduction

For the past 18 months, we have been building the comprehensive Terrorism Knowledge Base™ (TKB™). The TKB will ultimately contain all relevant knowledge (beginning with unclassified knowledge) about terrorist groups, their members, leaders, ideology, founders, sponsors, affiliations, facilities, locations, finances, capabilities, intentions, behaviors, tactics, and full descriptions of specific terrorist events. Led by world-class experts in terrorism supplied by SAIC, knowledge enterers have been building the TKB at the rate of up to 100 assertions per person-hour. The TKB knowledge entry tools are simple enough for unassisted domain experts to use. However, the knowledge entered has not been limited to simple database entries, but rather is stored as statements in mathematical logic, suitable for computer understanding and reasoning. In fact, the TKB has a growing complement of reasoning modules for integrating data and correlating observations, generating scenarios, answering questions and composing explanations. The TKB will be the first fully computer-understandable terrorist knowledge base capable of supporting computer reasoning to aid analysts.

2. Building the TKB in Cyc

The TKB is an augmentation of the existing Cyc® Knowledge Base (Cyc KB) (Lenat and Guha 1990; Guha 1995), which has been under intensive construction for the past 20 years. Most of the current content of the Cyc KB consists of general facts about many kinds of everyday objects and activities. It also contains "almanac-style" facts about individual countries, ethnic groups and organizations. Prior to launching the development of the TKB, the Cyc KB already had a substantial amount of knowledge relevant to terrorist activity. All of this knowledge is encoded in CycL, a language with the expressivity of higher-order predicate calculus.

The TKB effort has so far added to the Cyc KB knowledge of over two thousand terrorists, over seven hundred terrorist groups and over sixty-five hundred terrorist attacks. The representations of these individuals, groups and events are involved in over two hundred thousand TKB assertions such as "Xavier Djaffor participated in the Jihad from 1996 to 2000" and "Lashkar-e-Taiba is an Islamist terror group founded in 1990".

Subject-matter experts reading wire service reports, newspaper articles, *etc*., record information in the fields of the "Fact Entry Tool" (FET), which operates very much like a web form. The FET user looks up or creates some particular individual – a terrorist, terrorist attack, or terrorist group – and then enters information about that individual by filling in particular FET fields. In some cases, values for the fields can be selected by choosing them from a drop-down menu. The FET user can also type ordinary English into these fields, and the system will parse that English text into a representation in CycL. For example, if the user is prompted to enter the name of a person and a representation of the individual already exists in the TKB, the system parses the individual's name to the unique TKB knowledge structure that represents it. For example, "Ousama bin Laden" and "Usama bin Ladin" (among dozens of other strings) automatically parse to the CycL datastructure, `OsamaBinLaden`. If a representation of the individual does not already exist in the TKB, then the new individual is created.

Each assertion has its source (a web page, an expert, a newspaper article, *etc*.) associated with it and the source itself is represented as a first class object with assertions describing its name, date of issue, and its author or publisher (if relevant). Further, we keep track of which expert entered which data from each source, and when.

In addition to containing knowledge that has been entered through the FET, the TKB can also automatically access knowledge contained in any structured database by leveraging Cyc's Semantic Knowledge Source Integration (SKSI) technology (Masters and Güngördü,

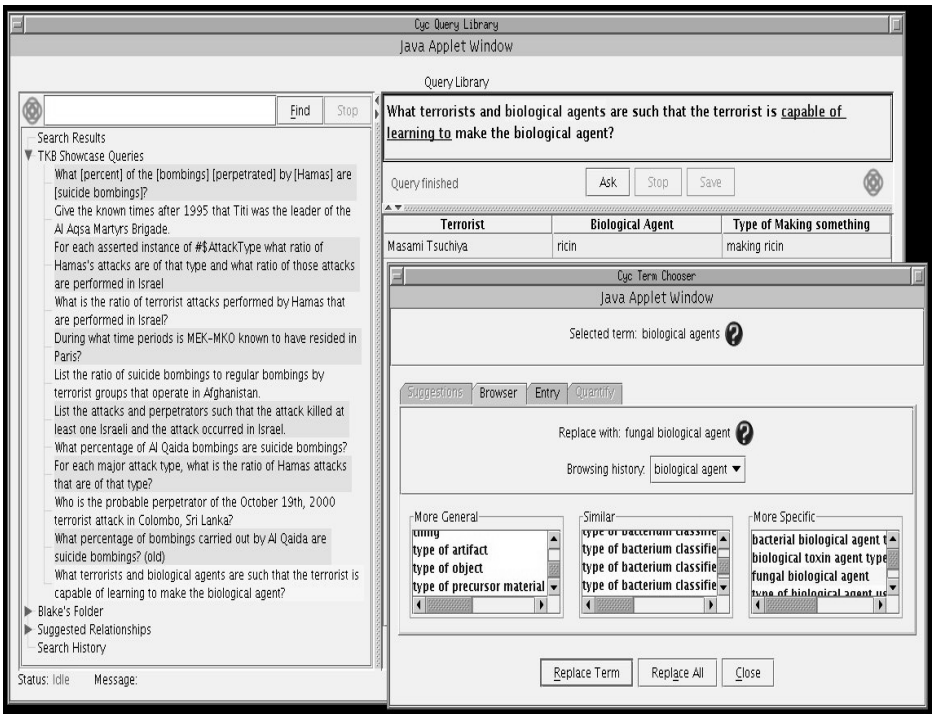

Figure 1. The Query Library

2003). Instead of composing distinct queries against various databases, analysts can query the TKB once, thereby accessing knowledge in the TKB as well as any of its *virtual* knowledge that actually resides in external semantically integrated databases or other sources.

3. Leveraging Cyc-style Inference

The Query Library is an interface to Cyc's inference capabilities (Figure 1). It enables users to ask queries against the TKB using query templates, which themselves are stored in the TKB. Query templates are saved with certain default values that users can replace by (1) entering English terms that Cyc parses into the appropriate CycL constants, or (2) using dedicated ontology browsing and search tools to find the desired term. The Query Library generates the appropriate English for the query so the user need never interact with the underlying CycL representations. For instance, a query to determine who is capable of learning to manufacture biological agents would be represented in CycL as follows:

```
(and
 (isa ?T Terrorist)
 (isa ?TOX BiologicalAgentType)
 (makingTypeOfType ?PROD ?TOX)
 (capableOfLearningTo ?T ?PROD performedBy))
```

However, the user would see the following:

"What terrorists and biological agents are such that the terrorist is capable of learning to make the biological agent?"

The underlined terms in this text are replaceable from within the Query Library Tool. The user, for example, could replace "biological agents" with a term that refers to a more specific collection (such as "fungal biological agents").

4. Future Work

Representing the relevant data using the FET is quite fast, but it still requires effort from subject-matter experts. A mechanism for automatically retrieving the relevant information from natural-language reports is a logical improvement in the process. Current work is focusing on improving natural language retrieval technologies for the TKB; we expect to achieve substantial gains by improving the TKB's integration with information extraction technologies (Appelt and Israel 1999).

The TKB will simplify the analyst's workflow by providing ready access to knowledge that currently must be inferred by humans from data scattered across disparate sources. Once the initial TKB is complete, and as the benefits of an up-to-date TKB become apparent, we expect that analysts themselves will be able to maintain its currency with a small amount of effort.

References

Appelt, D. E. and Israel, D. J. 1999. Introduction to Information Extraction, a tutorial prepared for IJCAI-99. http://www.ai.sri.com/~appelt/ie-tutorial/IJCAI99.pdf

Lenat, D. B. and Guha, R. V. 1990. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley.

Lenat, D. 1995. Steps to Sharing Knowledge, in *Toward Very Large Knowledge Bases*, Mars, N.J.I. (ed.) IOS Press.

Masters, J. and Güngördü, Z. 2003. Semantic Knowledge Source Integration: A Progress Report. *Integration of Knowledge Intensive Multi agent Systems*, Cambridge, Massachusetts, USA, 2003.