From Language to Knowledge

Boris Katz, Gary Borchardt, Sue Felshin & Howard Shrobe

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: A critical impediment to building intelligent computer systems is our inability to get information into the machine. The need for a “knowledge engineering priesthood” represents a “language to knowledge” bottleneck which precludes rapid encoding of available information.

Motivation: We believe that we can make it possible to populate new knowledge bases rapidly, accurately and completely, by allowing subject matter experts to bypass the “knowledge engineering priesthood” and to build knowledge bases directly, using normal means of communication such as spoken and written natural language and sketching.

Previous Work: A considerable amount of research has been directed at the problem of knowledge acquisition. Two of the better-known efforts in this area are the PROTEGE II system at Stanford KSL [3] and the EXPECT system at USC ISI [4]. However, it remains the case that the largest knowledge base construction efforts have been accomplished largely through manual entry of knowledge by knowledge engineers (e.g., the Cyc project [2]). Our effort builds on the START information access system, which was originally conceived as a mechanism for direct entry and retrieval of knowledge using simple English [1].

Approach: Our approach takes advantage of two interacting ideas: breaking the “language to knowledge” transformation into several stages, and incorporating human interaction wherever possible along the way. The principal stages are:

- *finding information* in available external resources,
- *standardizing it* into simple, canonical English, and
- *encoding it* as assertions in the target representation.

In addition, the second stage can be decomposed into substages of transforming an initial utterance into parse trees, transforming the parse trees into a logical form, and transforming the logical form into a canonical form.

Given this breakdown of the process, there are several ways in which human interaction can be exploited. During the first stage, finding information, the human can query the system to determine what knowledge is already in the knowledge base and what is available from external resources. During the second stage, standardizing the knowledge, the human can select sentences for entry into the knowledge base, re-express sentences as necessary for successful parsing by the system, define new terms, and confirm or reject the system’s standardizations of sentences. During the third stage, encoding the knowledge, the human can confirm or reject the system’s logical encodings of sentences. Throughout the process, the human can direct the system to iteratively repeat steps or substeps. Finally, once an encoding of new knowledge has been formed, the human can submit new queries to the system to assess the correctness of the knowledge.

This past year, we constructed a system that interactively translates English assertions into an editable, graphical representation of underlying content. This graphical representation serves as the logical form in our staged model of the language to knowledge transformation and also provides some elements of the transformation from logical form to canonical form. In particular, interactive parsing casts the assertions in simpler grammatical forms, while generation of the graphical representation provides us with a standardized treatment of verb argument structure, possessive relationships, class-subclass and class-instance relationships, referring expressions, sentential embedding, negation, modifiers, and quantifiers. Separately, the graphical nature of the output representation and its choice of simple representational constructs helps us obtain human feedback during the translation process.

Impact: By freeing up the knowledge acquisition process so that it can be performed by domain experts rather than knowledge engineers, we can vastly increase the human labor pool available for this task. In turn, this will spur a significant increase in knowledge base construction and the use of knowledge based systems.

Future Work: During the coming year, we plan to enhance our translator so that it provides additional assistance to users in defining terms and rephrasing input assertions. In addition, we plan to implement a capability that allows users to iteratively map terms to corresponding terms drawn from a limited target vocabulary.

Separately, we plan to construct a facility that can index knowledge structures such as the graphical representations generated by our translator, so that these structures can be retrieved in response to English queries. As part of this effort, we have started the construction of a web-based interface that enables humans to compose natural language annotations that describe information segments of many types.

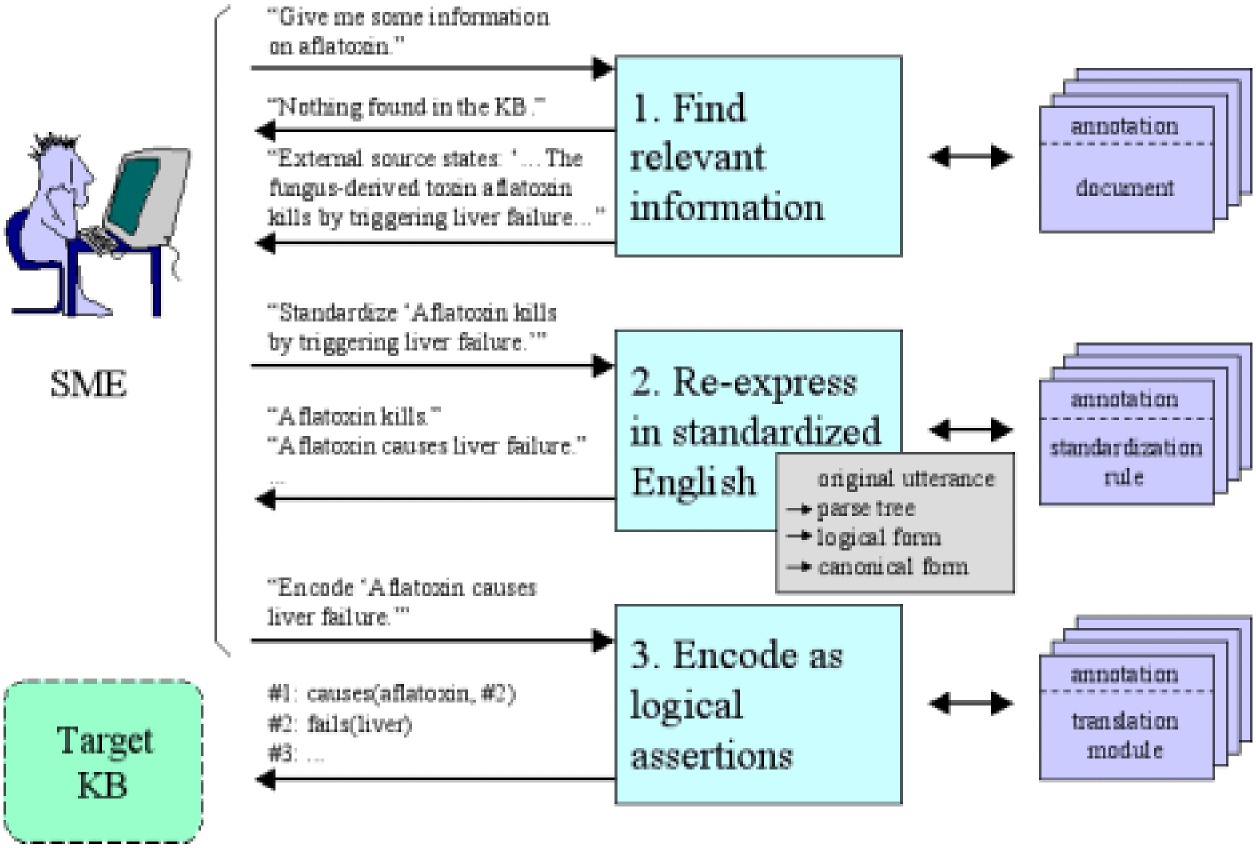
"Give me some information on aflatoxin."
"Nothing found in the KB."
"External source states: '…The fungus-derived toxin aflatoxin kills by triggering liver failure…'"
1. Find relevant information
annotation
document
SME
"Standardize 'Aflatoxin kills by triggering liver failure.'"
"Aflatoxin kills."
"Aflatoxin causes liver failure."
…
2. Re-express in standardized English
original utterance
→ parse tree
→ logical form
→ canonical form
annotation
standardization rule
"Encode 'Aflatoxin causes liver failure.'"
#1: causes(aflatoxin, #2)
#2: fails(liver)
#3: …
3. Encode as logical assertions
annotation
translation module
Target KB

Figure 1: Decomposition of the knowledge acquisition process into three successive stages, each supported by a combination of system functionality and human interaction.

Research Support: This research is funded by DARPA under contract number F30602-00-1-0545 and administered by the Air Force Research Laboratory.

References:

[1] B. Katz. Using English for indexing and retrieving. In P. H. Winston and S. A. Shellard, editors, *Artificial Intelligence at MIT: Expanding Frontiers*. MIT Press, 1990.

[2] D. B. Lenat, R. V. Guha, K. Pittman, and D. Pratt. Cyc: Towards programs with common sense. *Communications of the ACM*, 33(8):30–49, 1990.

[3] M. A. Musen, J. H. Gennari, H. Eriksson, S. W. Tu, and A. R. Puerta. PROTEGE II: Computer support for development of intelligent systems from libraries of components. Technical Report KSL-94-60, Knowledge Systems Laboratory, Stanford University, 1994.

[4] B. Swartout and Y. Gil. EXPECT: Explicit representations for flexible acquisition. In *Proceedings of the Ninth Knowledge Acquisition for Knowledge-Based Systems Workshop*, 1995.