Research Abstracts - 2007

The Rapid Story Annotation Workbench

Mark A. Finlayson & Patrick H. Winston

The Data Barrier for Computational Cognitive Modeling

Computational cognitive modeling seeks to reproduce psychological data gathered on human subjects with a computer model. Examples includes understanding human natural-language processing, reasoning by analogy, or the effects of culture on cognition. For it to be claimed that the computational model is actually modeling the experimental effects in question, the model must receive the "same" input as the human subjects. But because computers can't deal with the same level of complexity of stimuli that humans can, researchers need parallel datasets: one that is natural for people, another that is natural for computers, both containing (at least plausibly so) the same information. In computational models dealing with natural-language stimuli, this parallel data set takes the form of a set of stimuli pairs. One half of each pair is in natural language, to be read by human subjects; the other half is the same stimuli coded into some computer-parsable representation that purports to capture the relevant information for the model. One example of this type of dataset is the Karla the Hawk dataset (Gentner, 1993), which has been used to great effect in a number of cognitive modeling experiments.

The problem is that assembling these parallel data-sets is quite difficult: there is a "data-barrier" to comprehensive, natural-language-based computational cognitive modeling Writing the stories in natural English is easy, but translating them precisely and accurately into a computer-parsable semantic representation is not. Originally, researchers would hand-translate natural language stimuli into the representations appropriate for their models, but this manual translation is laborious, time-consuming, error-prone, and biased. In the past two decades, researchers have vigorously pursued automatic semantic parsing, that is, using a program to replace the manual annotator. While these researchers have made huge steps toward capabilities required for fully-automatic semantic parsing, they have not been able to bring language understanding to the level that is required for accurate and informative cognitive modeling

The Story Workbench Approach

We propose that a combination of the manual and the automatic approaches, fused with some sophisticated application engineering, will allow us to overcome the data barrier. We envision an application we call the *Rapid Story Annotation Workbench* which will combine statistical natural language techniques with user feedback. The idea is straightforward. The user inputs natural English (the stimulus that would be given to human subjects) and the application makes it's best guess of the appropriate formal representation, *a la* statistical techniques. While it is difficult for the computer to generate accurate semantic descriptions, it is fairly easy for it to check syntax and consistency, and to offer the user a constrained way of correcting them. Using such techniques, the user meets the application halfway, correcting ambiguities and problems and adding additional detail.

The workflow of the application is shown in *Figure 1*. First the user inputs text for interpretation. The application will use the latest in statistical natural language processing to make its best guess as to the appropriate semantic interpretation of the input (Step 2). The application then will display the results to the user, highlighting known or suspected problems and ambiguities (Step 3). Finally, the user will enter corrections or additional detail to achieve a correct semantic parse of the input (Step 4).

There are significant number of statistical techniques for syntactic and semantic parsing available for use in the workbench (for examples, see Manning, 1997). In particular, recent AI research has seen a boom in the abilities of statistical
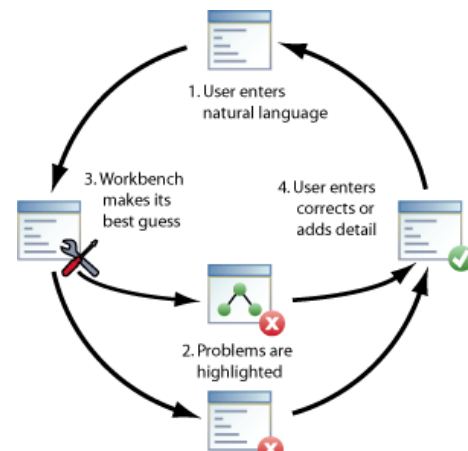

Figure 1: The proposed workflow of the Rapid Story Annotation Workbench. (1) The user enters natural language for interpretation. (2) The workbench makes its best guess as to the semantics of the text, and (3) Displays its interpretation and highlights ambiguities and problems for the user. (4) The user can run a variety of wizards or other user-friendly tools to correct the semantic representations or add detail to disambiguate the meaning of the text.

information-extraction techniques, such as entity-, date-, location-, and event-extraction from free text. Using these techniques, we have assembled a list of representations that we plan to include in the first version of the workbench, for application in modeling of cultural effects on cognition. These representations are intended to cover a large part of the semantics of cultural folktales and myths, and include (1) Syntactic structure & parts of speech, (2) Word sense, (3) Object identity, (4) Event identity, (5) Role relationships, (6) Causal relationships, (7) Temporal relationships, (8) Physical relationships, and (9) Verb Tense, Mood, & Aspect.

The Workbench will allow quick assembly of larger datasets than was ever before possible, and lead to quantum leap forward in kind and quality of computational cognitive experiments that are possible.

Current State of the Rapid Story Annotation Workbench

As of the time of this writing, the basic workbench framework has been completed. The application is being built on top of the Eclipse Integrated Development Environment. Eclipse has a plugin structure that allows capabilities to be extended without major refactorings of the code base. We are in the process of assembling the automatic description guessing algorithms and writing specific user interface components. So far we have incorporated a statistical natural language parser, and an interface to Wordnet for word-sense tagging. The next step is to implement an interface to the CYC knowledgebase and the Northwestern EA parser (Kuehne, 2004) for semantic tagging. We have also settled on formal frameworks for many of the semantic representations noted above, for example, the Allen time representation to express temporal relationships between events (Allen, 1984), and the RCC8 spatial representation for expressing physical relationships (Randall *et al.*, 1992) . A screenshot of the current state of the workbench is shown in *Figure 2*.
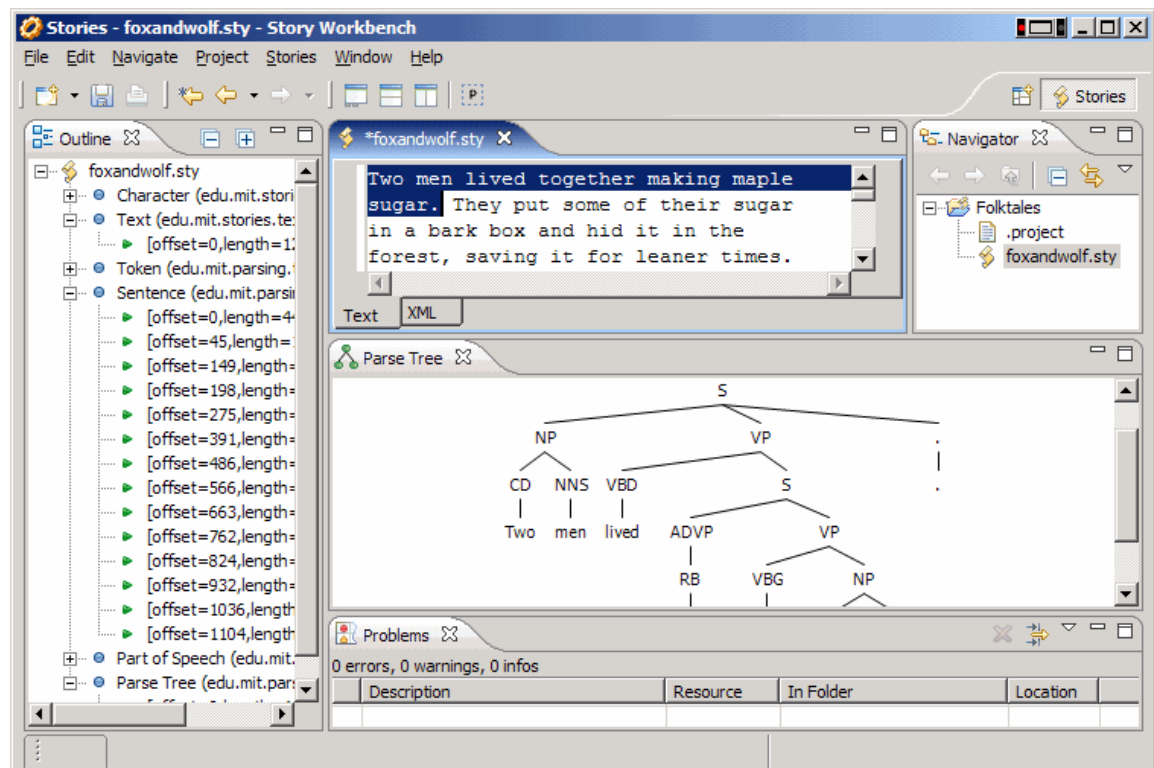

Figure 2: A screenshot of the current state of the workbench. In the *Outline* pane on the left, there is a snapshot of the current text, where each piece of the semantic description and it's location in the text is indicated. In the top-center pane is the Story Editor, which has an open story named `Fox and Wolf'. In the editor, a single sentence is selected, and the parse tree for this sentence is shown in the viewer directly below the Story Editor.

References

Allen, J. F. (1984). Towards a general theory of action and time. *Artificial Intelligence* **23**, 123-125.

Finlayson, M. A., & Winston, P. H. (2005). Intermediate features and informational-level constraint on analogical retrieval. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* **27**, 666-671. Stresa, Italy.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, **25**(4), 524-575.

Kuehne, S. E. (2004). *Understanding Natural Language Descriptions of Physical Phenomena.* Doctoral Thesis, Northwestern University. Evanston, IL.

Manning, C. D. and H. Schutze (1997). Foundations of Statistical Natural Language Processing. Cambridge, MA, MIT Press.

Randell, D. A., Z. Cui, and A.G. Cohn. (1992). A spatial logic based on regions and connection. In B. Nebel, W. Swartout, and C. Rich, editors, *Proceedings of the 3rd International Conference on Knowledge Representation and Reasoning*, 165–176, Los Allos, Morgan Kaufmann.