Current Research on "Smart Documents" in the Biological Knowledge Laboratory

by Robert P. Futrelle, Northeastern U., Boston futrelle@ccs.neu.edu

Abstract submitted 5/30/01 to the
Joint conference on Biological Research with Information Extraction BRIE'01 & Open-Access Publications OAP'01
to be held at the ISMB'01 Conference Tivoli Gardens, Copenhagen, Denmark July 26, 2001

Since its founding in 1989, the Biological Knowledge Laboratory (BKL) at Northeastern University has focused on knowledge-based approaches to the Biological literature, including work on both text and graphics (diagrams). This presentation will survey four areas of recent activity in the BKL and plans for the near future.

Work on large current biological journal corpora -- We have recently signed an agreement with a major publisher of biology journals to gain access to their full collection in HTML format. Initially this will comprise about 200 million words of text. The goal of our work on this corpus is to develop and use natural language processing and related techniques in machine learning and ontologies to generate knowledge representations for the contents of the papers that will in turn allow intelligent question answering, automated summarization, inter-document comparisons, etc. We are concerned that most past approaches to NLP have failed to deliver solid results so we are pursuing a new approach. Instead of focusing on grammars, the conventional rule-based approach, we will use a hybrid memory/pattern-based approach, inducing thousands of common patterns of expression from corpora and then analyzing specific instances of such patterns with regular expressions, which are fast pattern matchers. Instead of attempting to develop "deep" semantic representations, we will focus on near-surface-level transformations that directly link statements in the text to query patterns, since a major goal of these systems is to answer queries with great specificity. As our work on this corpus progresses we will apply the techniques we develop to the Medline corpus also.

www.bionlp.org -- Resources for Natural Language Processing of Biology Text -- I developed this site in February, 2001. It contains about 70 documents totalling 14MB. It will shortly have an associated mailing list, also hosted at Northeastern. I will explain its current status and seek feedback from meeting participants about the types of material that would be most useful.

The integrated nature of text and graphics in documents -- Diagrams and other graphics plays an important role in the majority of biological publications. But little has been done to characterize the role of graphics in such documents. Building on previous theories of text discourse, we have developed an approach to defining the semantics of graphics. This makes it possible to generate an integrated representation of text/graphics discourse. A major component of our theory is an integrated natural language / visual language lexicon that allows people to understand bimodal discourse. Another major component of text/graphics discourse is the "role of the reader". That is, the reader constantly exercises choices to shift his or her attention between the text and graphics while reading/viewing. We have demonstrated some computational approaches to the automation of building discourse structure at the level of syntax which is then followed by the construction of semantics via logical forms. A paper about this work was published and is available at http://www.dfki.de/~krueger/sg2001/schedule/futrelle.pdf

"Smart Documents" -- A paradigm for future research -- Our entire research effort can be summed up as work on "Smart Documents". We will describe our research program that is attempting to build both analysis tools and authoring tools for these content-based electronic documents of the future. The corpus-based NLP work just mentioned as well as all our work on Smart Documents is being situated in three-tiered systems, specifically, thin clients and object-based servers (all in Java) backed by an RDBMS (Oracle) adapted to deal with pattern structures. This offers fast, scalable performance that goes beyond conventional flat ascii files and ad-hoc data structures.

(This document is at: http://www.ccs.neu.edu/home/futrelle/brie2001/abstract-submitted.html)
The full presentation will be available later at http://www.ccs.neu.edu/home/futrelle/brie2001/