Final Project Report

Kenneth Baclawski

April 3, 1997

Part II

The point of this project was to find ways to use database management techniques to support biological research, specifically by developing techniques for analyzing, storing and querying biological research papers electronically. We created a knowledge model (ontology) by examining research articles on bacterial chemotaxis, focusing on the Materials and Methods sections, which give detailed descriptions of the ingredients, equipment and procedures used in an experiment. We then developed software for acquiring, storing and indexing knowledge from text. The knowledge frames were stored in an object-oriented database. Unlike previous methods of text analysis that rely on pattern matching directly on the text, we performed the pattern matching on text that has been tagged with part of speech tags and partially parsed using a robust parsing method. Biology uses a variety of highly specialized terms and phrases. Analyzing such text requires a knowledge of the properties of the biological lexicon, especially parts of speech and semantic classifications. Previous methods for lexicon development must be supervised by domain experts and are very labor intensive. We developed a method the unsupervised learning of the properties of a specialized lexicon. This method was applied to over 4 million words of text from the bacteriology literature.

Part III

The following papers and technical reports were based upon work supported by the grant:

[1] K. Baclawski and N. Fridman. M&M-Query: Database support for the annotation and retrieval of biological research articles. Technical Report NU-CCS-94-07, Northeastern University, College of Computer Science, 1994.

[2] K. Baclawski, R. Futrelle, N. Fridman, and M. Pescitelli. Database techniques for biological materials & methods. In *First Intern. Conf. Intell. Sys. Molecular Biology*, pages 21–28, 1993.

[3] K. Baclawski, R. Futrelle, C. Hafner, M. Pescitelli, N. Fridman, B. Li, and C. Zou. Data/knowledge bases for biological papers and techniques. In *Proc. Sympos. Adv. Data Management for the Scientist and Engineer*, pages 23–28, 1993.

[4] K. Baclawski, R. Futrelle, C. Hafner, M. Pescitelli, N. Fridman, B. Li, and C. Zou. M&M-Query: Materials & Methods knowledge base and query system. Technical Report NU-CCS-93-06, Northeastern University, College of Computer Science, 1993.

[5] K. Baclawski and B. Indurkhya. The notion of inheritance in object-oriented programming. *Comm. ACM*, 37:118–119, September 1994.

[6] K. Baclawski, D. Simovici, and W. White. A categorical approach to database semantics. *Math. Structures in Comp. Sci.*, 4:147–183, 1994.

[7] R. Futrelle. Interacting with large, persistent, object-oriented textbases: background and formulation. Technical report, Northeastern University, College of Computer Science, 1993.

[8] R. Futrelle and N. Fridman. Principles and tools for authoring knowledge-rich documents. In *DEXA 95 Workshop on Digital Libraries*, pages 357–362, London, U.K., 1995.

[9] R. Futrelle and X. Zhang. Large-scale persistent object systems for corpus linguistics and information retrieval. In *Digital Libraries '94*, pages 80–87, College Station, Texas, 1994.

[10] R. Futrelle, X. Zhang, and Y. Sekiya. Corpus linguistics for establishing the natural language content of digital library documents. In N. Adam, B. Bhargava, and Y. Yesha, editors, *Digital Libraries*, pages 165–180. Springer-Verlag, Berlin, 1995.

[11] J. Gray, P. Sundaresan, S. Englert, K. Baclawski, and P. Weinberger. Quickly generating billion record synthetic databases. In *Proc. ACM SIGMOD Conference*, pages 243–252, 1994.

[12] C. Hafner, K. Baclawski, R. Futrelle, N. Fridman, and S. Sampath. Creating a knowledge base of biological research papers. In *Proc. Second Intern. Conf. Intell. Sys. Molecular Biology*, pages 147–155, 1994.

[13] C. Hafner and N. Fridman. An ontology for substances and processes in Molecular Biology. In *Proc. First International Summer Institute in Cognitive Science*, Buffalo, NY, 1994. Center for Cognitive Science, SUNY at Buffalo.

[14] C. Hafner and N. Fridman. Ontological foundations for biology knowledge models. In *Proc. 4th International Conference on Intelligence Systems for Molecular Biology (ISMB-96)*, pages 78–87, Menlo Park, CA, 1996. AAAI Press.

[15] N. Noy and C. Hafner. The state of the art in ontology design: A comparative review. In *Proc. AAAI Spring Symposium on Ontological Engineering*. AAAI Press, March 1997.

[16] P. O'Neil, K. Baclawski, and F. Hsu. Designing computer networks to avoid partitioning. *Information Systems*, 18:343–348, 1993.

The following public domain software was developed based upon work supported by the project:

1. M&M Query System: supports text, topics and frame fill-in queries to the database of frames and documents.

2. Schema to English translator: produces a natural language description of a formal database schema.

3. Knowledge Definition Tool: allows one to annotate papers with the frames based on the underlying ontology.

4. Biology Materials & Methods Sublanguage grammar and lexicon.

5. MChart-F: an extension of the MChart parser for handling feature-based grammars and lexicons, accepting tagged or untagged input, and printing chart structures representing partially-parsed sentences in a symbolic form that can be easily understood by humans and programs.

6. Index Builder: builds indexes for words, objects and analysis results stored in persistent object-oriented databases to allow efficient access.

7. KWIC: displays the contexts of all occurrences of a given key word.

8. Word Classification: a statistical, totally unsupervised word classification program based on the Balanced Entropy Principle.

The public domain software listed above is available by anonymous ftp via the World Wide Web from the following home pages:

- `http://www.ccs.neu.edu/home/kenb`

- `http://www.ccs.neu.edu/home/futrelle`

- `http://www.ccs.neu.edu/home/hafner`

Part IV

	Senior Staff		Post-Doctorals		Graduate Students		Under-Graduates		Other Participants	
	Male	Fem.	Male	Fem.	Male	Fem.	Male	Fem.	Male	Fem.
A. Total, U.S. Citizens	3	1			1	1				
B. Total, Permanent Residents										
American Indian or Alaskan Native										
Asian										
Black, Not of Hispanic Origin										
Hispanic										
Pacific Islander										
White, Not of Hispanic Origin	3	1			1	1				
C. Total, Other Non-U.S. Citizens					2	4				
1. China					2	1				
2. Russia						1				
3. India						1				
4. Japan						1				
D. Total, All participants	3	1			3	5				
Disabled										