**Gary An, MD**
**Division of Trauma/Critical Care**
**Department of Surgery**
**Northwestern University Feinberg School of Medicine**
**201 East Huron, Galter 10-105**
**Chicago, IL 60611**
**docgca@gmail.com**

**and**

**Steven Lytinen, PhD**
**School of Computer Science, Telecommunications and Information Systems**
**DePaul University**
**243 S. Wabash, Room 645**
**Chicago, IL 60604**
**lytinen@cs.depaul.edu**

**Presentation Preference: Oral Presentation**

## Iterative Automated Named Entity Recognition to Bootstrap Biomedical Lexicon Generation

**Abstract:** Biocomplexity is manifest in the difficulty in characterizing many biomedical processes, with consequences for the development of therapeutic modalities. Furthermore, the scope of newly published research is overwhelming, impossible for an individual to effectively track and integrate. This suggests the need for collaborative efforts of the biomedical research community as a whole in order to integrate and concatenate this information. However, effective expression and communication of community-level knowledge is currently limited. Formal means of knowledge representation are required, both in terms of generating "knowledge-spaces" of existing information from the biomedical corpus, as well as being able to dynamically instantiate these "knowledge-spaces" in computational models to evaluate the consequences of their underlying hypotheses. We address this first aspect via automated text analysis and information extraction, an active area of development in bioinformatics. However, current methods are limited by the need to hand-build extensive dictionaries of cellular/molecular biology terms in order to apply information extraction methods. We propose an iterative method to evolve rule-based automated lexicon-generating software. We developed software to "read" the abstract list from the Shock 2005 meeting using rules for string recognition to pick out the nouns or noun groups which are likely to be relevant to the hypotheses asserted in the abstracts (i.e., the subjects and objects of the particular abstract). We used a previously hand collated list of the 85 abstracts. Rule refinement was carried out in an iterative fashion with the goal of progressively improving accuracy. After three iterations the Recall = 72% and Precision = 66%, approaching that of state-of-the-art named entity recognition software. Thus, we believe it is possible to bootstrap the development of biomedical lexicons using an iterative process of rule evolution. Embedded in a "wiki" process, this would utilize community knowledge and expertise, and provide updatable lexicon generation. This is a vital first step in the development of automated hypothesis extraction, and eventual dynamic knowledge representation networks using a distributed, open-source paradigm.