FIRST SHARED-TASK FOR CHALLENGES IN NATURAL LANGUAGE PROCESSING FOR CLINICAL DATA

Uzuner, O.[1]; Kohane, I.[2]; Szolovits, P.[3]
[1]University at Albany, SUNY, Albany, NY; [2]Brigham and Women's Hospital, Boston, MA; [3]CSAIL MIT, Cambridge, MA
Keywords: Natural Language Processing, Clinical Records, Data Release

Clinical records can be an important source of information for many studies. However, the information included in these documents is in the form of unstructured, ungrammatical, fragmented English text. Currently, tools for automatic linguistic processing (e.g., indexing, semantically interpreting, and searching) of these documents are very limited. Existing technologies for processing structured information such as databases and grammatical documents such as news articles have little utility for processing clinical records.

The lack of a standardized and large enough data set has been a major barrier to progress of NLP for clinical data. Efforts to solve NLP problems on clinical data have focused on individual, proprietary, private, small data sets. The resulting methods and tools are specific to their own data set, cannot be easily shared among researchers, nor can they be fairly compared to each other.

This year, i2b2 took major steps towards creating a standardized data set that will be the gold standard for various NLP tasks on clinical data. This data set consists of discharge summaries from Partners Healthcare and is distributed to all interested researchers for two major NLP challenges:

Task 1: automatic de-identification of clinical records, and
Task 2: identification of smoking status of patients.

The two challenges will be run as "shared tasks", i.e., all participants will address the same two challenges on the same data set; their results will be compared to the same gold standard. The results obtained from the systems of different participants will help us evaluate the relative strengths of different approaches to addressing the same problem. As a result, we will be able to identify the state of the art in NLP for these challenges so that future efforts can build on past experience.

Workshop. To complement the shared task efforts, i2b2 partnered with the American Medical Informatics Association to organize a workshop for challenges in NLP for clinical records. This workshop will meet on November 10, 2006 and will be the showcase venue for the participants of the shared task.

Funding. This research is supported by grant #U54-LM008748 on Informatics for Integrating Biology to the Bedside from National Institutes of Health National Center for Biomedical Computing.

E-mail: ouzuner@albany.edu