

Natural Language Processing: Can it Help Detect Cases and Characterize Outbreaks?

Matthew H. Samore, MD

*VA Salt Lake City Health Care System and the Departments of Internal Medicine
& Biomedical Informatics, University of Utah School of Medicine*

OBJECTIVE

To demonstrate how natural language processing (NLP) of clinical records can contribute to case detection and characterization in biosurveillance.

MODERATOR

Matthew Samore, MD, Professor of Internal Medicine and Biomedical Informatics, University of Utah School of Medicine.

PANELISTS

1. Brett South, MS: "Identifying Contextual Features to Improve the Performance of an Influenza-like Illness Text Classifier"
2. Wendy Chapman, PhD: "Identifying respiratory-related clinical conditions from ED reports with Topaz"
3. Sylvain DeLisle, MD, MBA: "Text processing to enhance detection of acute respiratory infections"
4. Adi Gundlapalli, MD, PhD, MS: "Using NLP on VA Electronic Medical Records to Facilitate Epidemiologic Case Investigations"

DESCRIPTION

Information described in clinical reports may be useful for biosurveillance but is locked in free-text format. NLP techniques aim to unlock that information by extracting and encoding relevant information for use by systems performing case detection, outbreak detection, and outbreak characterization. In this panel, we bring together four complementary research studies that address the question: Can NLP help detect cases and characterize outbreaks?" The overall goal is increasing specificity and enhancing biosurveillance.

Abstract 1: How much of a difference does accurate identification of contextual information make in detecting Influenza-like Illness from text? The first abstract describes an error analysis of an NLP-based ILI case detection application previously applied to the VA textual clinical records. The authors hypothesized that many of the false positives were due to misclassifications of negation, temporality, and note templating. To study the effect of such contextual information, they examined a subset of sentence strings from false positive cases and showed that spe-

cific contextual features and note templating contributed to false positive concepts.

Abstract 2: How does NLP compare to physicians in identifying concepts and their contextual properties? Having shown that contextual information is important for accurate case identification, we address the question: how well can NLP identify contextual properties of relevant clinical concepts? The second abstract evaluates a system called Topaz that was used to identify respiratory-related clinical conditions from Emergency Department reports. The authors compare Topaz to physicians in determining which of 55 respiratory-related conditions were presented acutely by patients in the study and show that Topaz performs similarly to physicians, a promising finding for NLP applications processing complex narratives.

Abstract 3: Does automatically extracted textual data improve case detection over structured data alone? Having demonstrated that an NLP application can perform quite well at determining which acute clinical conditions were experienced by patients in the ED, we next address a higher level question: How much does the information described in textual records contribute to case detection above what can be extracted from structured databases? The authors of abstract three measured whether clinical information extracted by NLP from clinical records enhances case detection based on structured data, such as measured fever and a new prescription for cough medication. They found that free-text processing of clinical notes brings information about disease symptoms that complements and enhances detection.

Abstract 4: Can NLP be used to improve case investigation? After detecting an outbreak, clinical records can be used to help characterize the outbreak. In abstract 4, the authors apply NLP to the task of extracting epidemiological variables, such as homelessness, infectious disease exposure, and important epidemiologic variables. They find that key investigation variables are available only in the free text and are amenable to extraction and identification by NLP.

STATEMENT OF PARTICIPATION

All authors have agreed to participate in a panel session.