A Dictionary-based Method for Detecting Anomalous Chief Complaint Text in Individual Records

Sara A. Taylor*[1] and Aaron Kite-Powell[2]

[1]Brigham Young University, Provo, UT, USA; [2]MIT Lincoln Laboratory, Lexington, MA, USA

Objective

To automate the detection of very unusual emergency department chief complaints based on a comparison between a trained dictionary of terms and the unstructured chief complaint field.

Introduction

The success of syndromic surveillance depends on the ability of the surveillance community to quickly and accurately recognize anomalous data. Current methods of anomaly detection focus on sets of syndromic categories and rely on a priori knowledge to map chief complaints to these general syndromic categories. As a result, the mapping scheme may miss key terms and phrases that have not previously been used. Furthermore, analysts do not have a good way of being alerted to these new terms in order to determine if they should be added to the syndromic mapping schema. We use a dynamic dictionary of terms to side-step the downfalls of a priori knowledge in this rapidly evolving field by alerting the analyst to rare and brand new words used in the chief complaint field.

Methods

We create a dictionary of all terms used in the chief complaint field of ED records from a training set comprised of state-wide emergency department data for sprint (March 1st through May 31st) of 2008, 2009, and 2010. This training set includes approximately 4.1 million records. The resulting dictionary contains approximately 50,000 words. We then create a second dictionary by restricting the full dictionary to the words that make up 97% of the words in the chief complaints in the training set, resulting in approximately 2,500 words. This is done so that we have a baseline of which words normally appear in a chief complaint text. Additionally, we create a “watched” list, which contains terms such as Anthrax, Dengue Fever, Measles, and Yellow Fever. We then flag an individual record in the test set (comprised of state-wide emergency department data for spring of 2011, approximately 1.6 million records) if the chief complaint field includes a word not found in the dictionary or if it contains a word on the “watched” list.

Results

Using the smaller dictionary our individual flagging method results in 4.59% of the records being flagged because of terms not found in the dictionary and 0.002% of the records being flagged from our watched list. Approximately 51% of the individual records flagged were not mapped into a specific syndromic category using a modified BioSense schema. When the new-word flagged records are broken down into day-by-day time-of-arrival windows, the mean rate of flagging (the number of flagged records on a particular day divided by the number of total records that day) is 4.51% or approximately 32 flagged records per day. This method flags records where the chief complaint field includes terms not found in the dictionary like “pinworms,” “parasite,” and “hydrocephalus” as well as catching the “watched” words like “anthrax.” However, this method results in a lot of false alarms because of spelling and other typographical errors which we believe can be reduced by better preprocessing.

Conclusions

Using a dictionary of frequently used terms and a list of “watched” terms to compare to the chief complaint field of emergency department records allows the analyst to be alerted to new and rare terms that they might have otherwise missed in their searching of the records. For instance, during the 2009 influence pandemic “H1N1” in chief complaints may not have been found unless it’s included with the other symptoms that bin it into a relevant syndrome category that is observed by an analyst, and therefore may go un-noticed for a period of time.

In future work, we plan on including an analyst in the loop and allowing them to decide if the word that was flagged as being anomalous was truly a new or rare term. We believe that this will increase the percentage of genuine alarms and will allow for the dynamic dictionary to be easily maintained.

Keywords

Syndromic Surveillance; chief complaint text; anomaly detection; dictionary-based

Acknowledgments

This work is sponsored by the Defense Threat Reduction Agency under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, recommendations and conclusions are those of the authors and are not necessarily endorsed by the United States Government

*Sara A. Taylor
E-mail: sara@ehlert.org