Development of a Custom Spell-Checker for Emergency Department Data

Sophie Rand, Ramona Lall

Bureau of Communicable Diseases, NYC DOHMH, Long Island City, New York, United States

Objective

To share progress on a custom spell-checker for emergency department chief complaint free-text data and demonstrate a spell-checker validation Shiny application.

Introduction

Emergency department (ED) syndromic surveillance relies on a chief complaint, which is often a free-text field, and may contain misspelled words, syntactic errors, and healthcare-specific and/or facility-specific abbreviations. Cleaning of the chief complaint field may improve syndrome capture sensitivity and reduce misclassification of syndromes. We are building a spell-checker, customized with language found in ED corpora, as our first step in cleaning our chief complaint field. This exercise would elucidate the value of pre-processing text and would lend itself to future work using natural language processing (NLP) techniques, such as topic modeling. Such a tool could be extensible to other datasets that contain free-text fields, including electronic reportable disease lab and case reporting.

Methods

Chief complaints may contain words that are incorrect if they are misspelled (e.g.,"patient has herpertension"), or, if the word yields a syntactically incorrect phrase (e.g., the word "huts" in the phrase: "my toe huts").

We are developing a spell-checker tool for chief complaint text using the R and Python programming languages. The first stage in the development of the spell- checker is the identifying and handling of misspellings; future work will address syntactic errors. Known abbreviations are identified using regular expressions, and unknown abbreviations are addressed by the spell-checker. The spell checker performs 4 steps on chief complaint data: identification of misspellings, generation of a substitute candidate word list, word sense disambiguation to identify replacement word, and replacement of the misspelled word, based on methods found in the literature [1]. As the spell-checker requires a dictionary of correctly spelled, healthcare-specific terms including all terms that would appear in an ED corpus, we used vocabularies from the Unified Medical Language System, ED-specific terminology, and domain expert user input. Dictionary construction, misspelling identification algorithms, and word list generation algorithms are in the development stage.

Simultaneously, we are building an R Shiny interactive web application for syndromic surveillance analysts to manually correct a subset of misspelled words, which we will use to validate and evaluate the performance of the spell-checker tool.

References

1. Tolentino HD, Matters MD, Walop W, et al. 2007. A UMLS-based spell checker for natural language processing in vaccine safety [PubMed]. *BMC Med Inform Decis Mak*. 7(1). doi:https://doi.org/10.1186/1472-6947-7-3. PubMed

Results

Project still in development phase.

Conclusions

The audience will learn about important considerations for developing a spell-checker, including those for data structure of a dictionary and algorithms for identification of misplaced words and identification of candidate replacement words. We will demonstrate our word list generation algorithm and the Shiny application which uses these words for spell-checker validation. We will share relevant code; after our presentation, audience members should able to apply code and lessons to their own projects and/or to collaborate with the NYC Department of Health and Mental Hygiene.