Evaluation of Preprocessing Techniques For Chief Complaint Classification

Jagan Dara, M.S., Wendy W. Chapman, Ph.D
Center for Biomedical Informatics, School of Medicine, University of Pittsburgh

OBJECTIVE

To determine whether preprocessing chief complaints before automatically classifying them into syndromic categories improves classification performance.

BACKGROUND

The Real-time Outbreak and Disease Surveillance (RODS) system collects chief complaints as free text and uses a naïve Bayesian classifier called CoCo to classify the complaints into syndromic categories [1]. CoCo 3.0 has been trained on 28,990 manually classified chief complaints. The free text chief complaints are challenging to work with, due to problems caused by linguistic variations such as synonyms, abbreviations, acronyms, truncations, concatenations, misspellings and typographic errors [2]. Failure to correct these word variations may result in missed cases, thereby decreasing sensitivity of detection.

METHODS

We developed a preprocessing module that 1) replaces abbreviations and truncations with expanded forms, 2) corrects misspellings, and 3) removes words that do not have clinical meaning. We used a test set of 10,161 chief complaints not previously involved in CoCo's training to measure the proportion of chief complaints changed by the preprocessor. We counted the number of unique words in the training set for CoCo 3.0 prior to and post preprocessing, and computed the proportion of words changed by the preprocessor that already existed in CoCo's training set. We measured CoCo's syndromic classification performance on the 10,161 chief complaints, with and without the preprocessing module. Reference standard classifications for the chief complaints were generated by consensus by a physician board-certified in internal medicine and infectious diseases with 30 years of experience and two emergency department ICD diagnosis coders, who used case definitions we developed to classify the chief complaints into any of seven syndromes. We calculated sensitivity and specificity of classification. In addition, for syndromes whose sensitivity decreased with the preprocessing module, we performed an error analysis on the chief complaints whose classification changed from correct without preprocessing to incorrect with preprocessing.

RESULTS

The preprocessor changed 59% of the chief complaints and decreased the number of unique words in the training set from 2,775 to 2,308. All the words changed in the test set by the preprocessor existed in the original training set. Table 1 shows the sensitivity and specificity of classification for each syndrome prior to and after preprocessing.

	Sensitivity		Specificity	
Syndrome	**PP**	**AP**	**PP**	**AP**
Botulinic	0.58	0.53	1.00	1.00
Constitutional	0.47	0.57	0.99	0.98
Gastrointestinal	0.70	0.67	0.99	0.99
Hemorrhagic	0.65	0.68	0.99	0.99
Neurological	0.62	0.64	0.97	0.97
Other	0.97	0.97	0.86	0.88
Rash	0.77	0.81	1.00	1.00
Respiratory	0.79	0.83	0.98	0.98

Table 1 – Sensitivity and Specificity of classification for each syndrome, prior to (**PP**) and after preprocessing (**AP**)

Sensitivities for most of the syndromes slightly increased. However, sensitivity for Gastrointestinal and Botulinic syndromes decreased after preprocessing. An error analysis showed that only a few (4/73) mistakes were directly due to the preprocessor. The majority of the remaining errors (60/73) involved CoCo's classification of complaints with multiple problems. CoCo currently selects the classification with the highest probability, even though a complaint may have several correct classifications. Sixty classifications that were originally Botulinic or Gastrointestinal were changed after preprocessing but were still correct. For example, initially, CoCo classified "abd pain blood in urine" as Gastrointestinal, but after preprocessing, CoCo classified the complaint as Hemorrhagic, which is also correct.

CONCLUSIONS

The preprocessing only slightly increased CoCo's sensitivity. One possibility is that because CoCo is a statistical classifier, CoCo was already trained to correctly classify the misspellings and abbreviations in the test set. We plan to test the preprocessor on a non-statistical system, such as a keyword search algorithm. In the future, we will incorporate into the preprocessor a module for splitting multiple problems before classification.

REFERENCES

[1] Olszewski RT. Bayesian classification of triage diagnoses for the early detection of epidemics. In: Recent Advances in Artificial Intelligence: Proceedings of the Sixteenth International FLAIRS Conference; 2003: AAAI Press; 2003. p. 412-416.

[2] Shapiro AR. Taming Variability in Free Test: Application to Health Surveillance. MMWR Vol.53 Sept 2004:95-100.

Further Information: Jagan Dara, jdara@cbmi.pitt.edu