Identifying Contextual Features to Improve the Performance of an Influenza-like Illness text Classifier

Brett R. South, MS[1,2,6], Shuying Shen MStat[1,2,6], Sylvain Delisle, MD,MBA[4], Trish Perl, MD,MSc[5], Matthew H. Samore, MD[1,2,6], Adi V. Gundlapalli, MD,PhD,MS[1,2,6], Wendy W. Chapman, PhD[3]

[1]VA Salt Lake City Health Care System, [2]Department of Internal Medicine, University of Utah School of Medicine, [3]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, [4]VA Maryland Health Care System and University of Maryland School of Medicine, [5]Johns Hopkins Medical Institutions and University, [6]Biomedical Informatics, University of Utah School of Medicine

OBJECTIVE

To understand the types of false positive cases identified by an Influenza-like illness (ILI) text classifier by measuring the prevalence of ILI-related concepts that are negated, hypothetical, include explicit mention of temporality, experienced by someone other than the patient, or described in templated text that is difficult to process.

BACKGROUND

Automated text classifiers rely on extraction of clinical concepts based on symptoms, problems, or findings from electronic note documents. False positive extractions may be due to concepts in the text being assigned to the patient when in reality the concepts were described as negated, hypothetical, explicitly mention duration, or experienced by someone other than the patient. Accurate identification of concepts can also be affected by peculiarities associated with electronic documents generated by the combination of free-text provider input and templated note structures used by Electronic Medical Record (EMR) systems.

METHODS

We manually reviewed sentences from patients who were incorrectly classified as patients with ILI using an existing automatic text classifier and annotated the ILI-related concepts in the sentences. We measured the proportion of relevant concepts modified by the properties described above.

We randomly sampled 25,403 electronic note documents from the VA EMR of two VA Medical Centers between October 2003 and March 2004. The note corpus included the full text of note documents commonly used for automated biosurveillance purposes, including chief complaint strings, emergency department notes, and nurse triage notes. Concepts used by our text classifier for case detection included the following eight ILI symptoms: cough, fever, chills, night sweats, pleuritic chest pain, myalgia, sore throat, or headache. Symptoms were mapped to a standard vocabulary using the UMLS Metathesaurus[1]. We first ran the unmodified version of a negation algorithm called NegEx[2] coupled with mapped ILI concepts on the note corpus and evaluated text classifier statistical performance. Presence of two or more unique non-negated concepts denoted ILI. We selected a random sample of sentence strings from false positive cases identified by the text classifier containing any one of the ILI concepts. Using an open source tool called Knowtator[3] we conducted a manual annotation of these strings. Two reviewers annotated the same document set with a third reviewer completing a blinded arbitration to arrive at a consensus set.

RESULTS

Sensitivity and PPV of the unmodified version of the text classifier applied to surveillance document sources was 75% and 27% with 569(4%) false positive cases. We reviewed 1,000 sentence strings randomly sampled from a total of 1,467 strings associated with 429 false positive cases. Inter-annotator agreement for manual annotation of concepts was 0.98. A total of 1,468 ILI concepts were identified in selected sentence strings. The prevalence of the relevant properties and note templating in sentence strings was 246(17%) negation, 21(1.4%) hypothetical, 111(7.6%) duration >7 days, and note templating including 405(28%) pick-list of signs and symptoms, and 94(6.4%) clinician or medication instructions. Only one out of a thousand strings mentioned an experiencer other than the patient.

CONCLUSION

These findings suggest that these properties are quite prevalent in false positive ILI cases. Modifying our ILI text classifier to address relevant features and note templating could potentially improve text classifier precision.

REFERENCES

1. http://umlsks.nlm.nih.gov, 2007(Version 5.0).
2. Chapman WW, et.al. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301-10.
3. http://knowtator.sourceforge.net

Further information: brett.south@hsc.utah.edu