Analysis of False Positive Errors of an Acute Respiratory Infection Text Classifier due to Contextual Features

Brett R. South, MS[1,2,5], Shuying Shen, MStat[1,2,5],
Wendy W. Chapman, PhD[3], Sylvain Delisle, MD, MBA[4], Matthew H. Samore, MD[1,2,5],
Adi V. Gundlapalli, MD, PhD, MS[1,2,5]
[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,
[4]VA Maryland Health Care System and University of Maryland School of Medicine,
[5]Department of Biomedical Informatics, University of Utah School of Medicine

Abstract

Text classifiers have been used for biosurveillance tasks to identify patients with diseases or conditions of interest. When compared to a clinical reference standard of 280 cases of Acute Respiratory Infection (ARI), a text classifier consisting of simple rules and NegEx plus string matching for specific concepts of interest produced 569 (4%) false positive (FP) cases. Using instance level manual annotation we estimate the prevalence of contextual attributes and error types leading to FP cases. Errors were due to (1) Deletion errors from abbreviations, spelling mistakes and missing synonyms (57%); (2) Insertion errors from templated document structures such as check boxes, and lists of signs and symptoms (36%) and; (3) Substitution errors from irrelevant concepts and alternate meanings for the same word (6%). We demonstrate that specific concept attributes contribute to false positive cases. These results will inform modifications and adaptations to improve text classifier performance.

Introduction

The goal of biosurveillance is timely case detection and investigation of potential disease outbreaks by hospitals and public health authorities. This is of practical significance for clinical care and for instituting control strategies to prevent transmission of disease within the population at risk. In settings where electronic clinical documents are available, Natural Language Processing (NLP) can be used to develop automated information extraction (IE) methods to extract and classify clinical information useful for biosurveillance. Extracted information can then be used to reduce the workload required for case finding and investigation. This assumes greater importance when dealing with large numbers of patient records, limited resources and an urgent need to identify patients of interest.

In most instances, ill patients presenting to the hospital with classical symptoms are suspected of having pandemic influenza and can be easily tracked for surveillance purposes. However, as the number of such patients increases or alternatively, patients admitted for other diagnoses subsequently exhibit symptoms of influenza, manually tracking these patients for outbreak investigation and isolation poses a significant challenge. In these circumstances, it would be beneficial to have an automated system to identify patients with symptoms of pandemic influenza. Developing a clinical informatics solution using automated IE methods has the potential to improve patient care and reduce the workload for those involved in surveillance.

Depending on the goals of a surveillance system, simple or complex IE and text classification techniques may be used. Simple text classifiers rely on accurate extraction of semantic concepts representing symptoms, problems and findings from clinical free text documents. Lists of semantic concepts can be expanded using the UMLS Methathesaurus to identify synonyms and term variants[1]. Concept lists are frequently coupled with a negation algorithm[2-4] and rules are applied to further assess what conditions the patient is actually experiencing and those conditions that are absent.

Text classifier accuracy can be improved by reducing extraction of concepts that are in reality negated, hypothetical, temporaly unrelated to the event, or experienced by someone other than the patient[5]. Correctly identifying contextual attributes of signs or symptoms is important to determine whether the condition is present or absent in the patient. Accurate concept extraction can also be affected by peculiarities associated with electronic documents generated by the combination of free text provider input and templated clinical note structures characteristically used by Electronic Medical Record (EMR) systems.

Background

Previous efforts that have applied IE methods to free text clinical documents for the purpose of biosurveillance have primarily focused on extracting concepts of interest from a limited set of *data* *sources*, such as those that include chief complaint text, emergency department visit notes, and nurse triage notes. In settings where a full EMR is available there are potential opportunities for the practical application of information extraction methods on *all electronic free text data sources*. Characteristics of EMR systems that are particularly useful for biosurveillance purposes include a rich source of structured data elements coded with standard vocabularies and unstructured data elements in form of free text clinical notes. Information sources that are both timely and can be readily and accurately extracted from encounter notes and made available for case finding and investigation purposes are particularly important for biosurveillance efforts.

Using Acute Respiratory Infection (ARI) as an example, this pilot study was undertaken to demonstrate attributes of concepts that result in false positive (FP) cases when applying a text classifier to a corpus of electronic clinical documents. To do so, we applied manual annotation methods to conduct an instance level error analysis with the goal of reducing extraction of concepts that contribute to FP cases.

Setting

This study was carried out using data and resources from two large Veterans Health Administration (VHA) healthcare facilities in the United States that use an integrated paperless EMR system for patient care. These two facilities provide care for nearly 90,000 patients with an average of over one million yearly outpatient encounters producing approximately three million electronic clinical notes per year.

Methods

Study Population, Case Definition, and Reference Standard

For this study 76,500 electronic medical notes from a random sample of 15,377 patient encounters at the two healthcare facilities between October 2003 and March 2004 were reviewed manually to identify patients with clinical features of ARI and generate a clinical reference standard. A patient was considered positive for ARI if: (1) the patient had a positive influenza culture or influenza antigen or (2) any two of the following symptoms were present for ≤7 days duration: cough, fever or chills or night sweats, pleuritic chest pain, myalgia, sore throat, headache; and (3) illness was not attributable to non-infectious etiology.

Text Classifier

For this pilot study, we were interested in applying the text-classifier to only those documents sources commonly used for biosurveillance. A rules based text classifier consisting of the unmodified NegEx version 2[6] plus string matching for concepts[7], was applied to a corpus of 10,439 electronic notes commonly used for automated biosurveillance purposes. This documents set included chief complaint strings, emergency department, and nursing notes. Concepts used by the text classifier included the following eight symptoms: cough, fever, chills, night sweats, pleuritic chest pain, myalgia, sore throat, or headache. Using the UMLS Metathesaurus[1], a final list of 186 concepts was assembled by mapping the symptoms from the case definition to a standard vocabulary. The final concept list included other clinically relevant terms identified from chart review efforts used to create the clinical reference standard (Table 1).

Table 1. Concepts related to Acute Respiratory Infection

Semantic concept	Number of synonyms, term variants, abbreviations
Cough	13
Fever	39
Chills	14
Night sweats	12
Pleuritic chest pain	14
Myalgia	29
Sore throat	35
Headache	30

The output from the text classifier included sentence strings in which ARI concepts were identified, cases of ARI along with sentence strings, concept(s), concept unique identifier (CUI), negation terms, status, and span of ARI related concepts and negation terms. *Presence of two or more unique non-negated concepts in the same clinical note denoted cases of ARI.* The statistical performance of the text classifier was determined by comparing these results to the clinical reference standard.

Our first objective was to conduct an instance level annotation of false positive cases at the concept and concept attribute level. False positive (FP) cases were identified based on discrepancies between the text classifier output and the clinical reference standard. A random sample of 1,000 sentence strings associated with FP cases were selected for manual annotation by human reviewers.

Manual Annotation: Tasks and Tools

An annotation schema was developed and implemented using an open source Protégé[8] plug-in tool called Knowtator[9]. All ARI concepts and attributes found in a random sample of 1,000 sentence strings were manually annotated identifying concept attributes of (1) Negation (affirmed, negated, hypothetical); (2) Duration of symptoms (≤7 days, >7days, unknown); (3) Experiencer (patient, family member, other); (4) Templating (instructions, signs/symptoms, other). Two reviewers annotated all 1,000 sentence strings and a third reviewer arbitrated disagreements. Annotators were only provided the pre-processed output sentence string in which ARI concepts were identified by the text classifier.

We estimate annotator performance on annotation tasks based on inter-annotator agreement (IAA) as described by Hripcsak[10] and Roberts[11] and calculated using the following formula:

IAA = (matches)/(matches+nonmatches).

An annotation guideline was created for this task and used for all manual annotation efforts. Based on methods described by Chapman[12], annotators first trained on a smaller set of documents to achieve an acceptable IAA using the annotation guideline and Knowtator annotation schema prior to completing the string level annotation tasks for FP cases.

In addition to a more traditional error analysis we were also interested in applying instance level manual annotation to identify and categorize types of error into the following categories: (1) *Substitution* error which occurs in situations where the concept is incorrect; (2) *Insertion* error which occurs where the concept is spurious; (3) *Deletion* error which occurs where the concept is missing. These types of classifications help to understand and characterize sources of error at the concept and concept attribute levels.

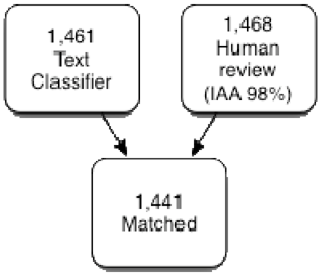
Our second objective for this study was to understand and characterize false positive (FP) cases at the concept and concept attribute level. To achieve this objective, we compared the output of concepts and attributes from the text classifier with annotation of sentence strings for FP cases.

Results

Of the 15,377 patient encounters at the two healthcare facilities, a total of 280 patients with a diagnosis of ARI were identified as the clinical reference standard by manual chart review (prevalence of the clinical condition in a random sample of patients was 1.8%). The recall (sensitivity) and precision (positive predictive value PPV) of the text classifier applied to surveillance document sources as described above was 75% and 27% respectively. The text classifier identified a total of 569 (4%) false positive cases with included concepts and concept attributes.

One thousand sentence strings randomly sampled from a total of 9,142 sentence strings, representing 1,467 notes associated with the 569 false positive cases were reviewed by two annotators. Inter-annotator agreement for manual annotation of concepts was 0.98. The distribution of concepts identified by the text classifier and manual annotation is shown in Figure 1.

Figure 1. Concepts identified by the text classifier and manual annotation. IAA = Inter-Annotator Agreement



Contextual attributes identified by manual annotation

A total of 1,468 ARI concepts were identified in selected sentence strings. The prevalence of the relevant properties and note templating in sentence strings is shown in Table 2.

Table 2. Concepts identified by manual annotation. IAA = Inter-Annotator Agreement

Attribute (IAA)	Value	Count (%)
Negation (93%)	affirmed	884 (60%)
	hypothetical	157 (11%)
	negated	427 (29%)
Duration (92%)	<=7 days	149 (10%)
	> 7 days	112 (8%)
	unknown	1207 (82%)
Templating (93%)	Signs and symptoms	405 (28%)
	Instructions	94 (6%)
	Free text only	968 (66%)

Among the concepts annotated in false positive cases, a majority (60%) were affirmed, while 29% were negated. Suggesting problems with negation processing. With regard to duration of symptoms, mentions were not explicit, resulting in a majority being of unknown duration (82%). Templated document structures represented a significant feature of annotated sentence strings (34%).

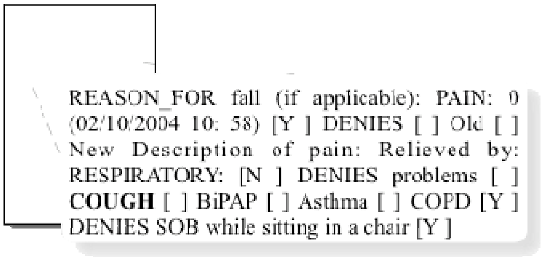
Discrepancies at the concept level

In addition to the discrepancies noted above due to contextual features, three types of discrepancies between text-classifier and human annotation of FP cases were noted at the concept level.

1) *Deletion errors* which occurred in situations where abbreviations, spelling mistakes and synonyms were missing from the concept list used by the classifier. These were identified by manual annotation and missed by the text classifier. Examples included abbreviations such as (HA, HA's, c, f, H/A, Ha's, ST), misspellings (shaking cills, fevrc), or synonyms that were missing from the original concept list (irritated throat, scratchy throat, myalgias).

2) *Insertion errors* which occurred in situations where concepts were identified by the classifier but not identified by human reviewer. Templated document structures including check boxes, long lists of signs or symptoms, or past medical history information accounted for the majority of these errors (Figure 2). In these types of strings there were also occurrences where negation is implied but not completed in the templated section due to unfilled check boxes.

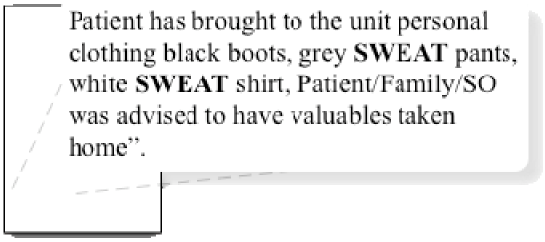
Figure 2. Example of templated pick list



REASON_FOR fall (if applicable): PAIN: 0 (02/10/2004 10: 58) [Y] DENIES [ ] Old [ ] New Description of pain: Relieved by: RESPIRATORY: [N] DENIES problems [ ] **COUGH** [ ] BiPAP [ ] Asthma [ ] COPD [Y] DENIES SOB while sitting in a chair [Y]

3) *Substitution errors* that occurred where irrelevant concepts were found due to an alternate meaning of the same word or a concept was present but out of context for this clinical use case (Figure 3).

Figure 3. Example of an out of context concept



Patient has brought to the unit personal clothing black boots, grey **SWEAT** pants, white **SWEAT** shirt, Patient/Family/SO was advised to have valuables taken home".

The discrepancy arises from an alternative meaning for the word "SWEAT"- which was found in our list of UMLS concepts, whereas in this sentence it refers to a type of clothing.

These particular types of discrepancies suggest problems with negation detection, identification of contextual features, and templated note structures that introduce processing error and contribute to false positive cases.

Limitations

We only looked at one syndrome of interest (ARI) for the preliminary results presented in this paper. We are currently testing these methods on other disease categories. The original reference standard of ARI cases was determined by manual review of charts first by a non-physician and then by a panel of physicians. It is possible that we missed some cases of ARI using this approach. Inter-annotator agreement may be over estimated since we did not test human annotation tasks without machine pre-processing. Though we provide examples of discrepancies at the concept level between human annotation and machine processing, additional review is necessary to quantify these error types. Improving identification of contextual features, including negation processing, and dealing with templated note structures that include unchecked check boxes may improve precision at the concept level reducing the number of false positive cases.

Conclusions

The performance of our text classifier in identifying cases of ARI was less than optimal and generates false positive cases. To modify and improve our text classifier, it is important to understand how these FP cases are generated at the concept and concept attribute level. Our pilot study has shown that such a review and error analyses can yield important information that can be used to further refine classifier performance.

Specific attributes such as ambiguities in negation of concepts and in determining the duration of symptoms lead to FP cases. Another important factor leading to FP cases is document templating that frequently occurs in electronic medical records. This refers to pre-defined sets of signs, symptoms or instructions that are associated with check boxes; thus they facilitate rapid assessment and documentation. However, leaving check boxes unchecked may lead to ambiguities in machine processing. Particularly in situations where interpretation is necessary to determine if items were simply unchecked because that item was not present or was not even asked of the patient. These properties of the text classifier may be amenable to improvements based on results of the error analyses and methods described by Denny et al[13].

Clinician notes represent a large proportion of patient information in the VHA electronic medical records system. NLP techniques provide a means of utilizing clinical documents as an additional source of data for surveillance. Moreover, utilizing NLP methods for potential case detection and epidemiologic investigation could potentially reduce the amount of time required for outbreak investigation. Informatics data sources such as clinical free text data have the potential to provide novel information not available in structured format that can be used to enhance case detection methods.

The results of this pilot study inform future efforts to improve precision by identifying contextual features and processing of templated note structures. This work also demonstrates one method of manually annotating the output from a text classifier and carrying out an error analysis at the concept level. Ongoing and future work includes further adaptation based on the error analyses reported in this paper, more detailed analyses of false negative cases for ARI, and extending these methods to other diseases and conditions of interest.

Acknowledgements

This study was supported using resources and facilities at the VA Salt Lake City Health Care System and VA Maryland Health Care System. The authors also gratefully acknowledge funding from the CDC for the BioSense Evaluation Grant (Johns Hopkins University) and Center of Excellence in Public Health Informatics (University of Utah).

References

1. UMLSKS (Unified Medical Language System Knowledge SourceServer) at http://umlsks.nlm.nih.gov,.2007 (Version 5.0).
2. Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. J Am Med Inform Assoc2001 Nov-Dec;8(6):598-609.
3. Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, et al. A controlled trial of automated classification of negation from clinical notes. BMC Med Inform Decis Mak 2005;5(1):13.
4. Huang Y, Lowe HJ. A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. J Am Med Inform Assoc 2007 Feb 28.
5. Goldin I, Chapman WW. Learning to detect negation with "not" in medical texts. Proceedings of the Workshop on Text Analysis and Search for Bioinformatics at the 26th Annual International ACM SIGIR Conference (SIGIR-2003). Eds. Eric Brown, William Hersh and Alfonso Valencia.
6. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform; 2001. p. 301-10.
7. South BR, Chapman W, Delisle S, et al. Optimizing A Syndromic Surveillance Text Classifier for Influenza-like Illness: Does Document Source Matter? AMIA Annu Symp Proc. 2008:692-6.
8. Ogren PV, Savova G, Buntrock JD, Chute CG. Building and evaluating annotated corpora for medical NLP systems. AMIA Annu Symp Proc. 2006:1050.
9. Musen MA, Gennari JH, Eriksson H, Tu SW, Puerta AR. PROTEGE-II: computer support for development of intelligent systems from libraries of components. Medinfo. 1995;8 Pt 1:766-70.
10. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform. 2002 Apr;35(2):99-110.
11. Roberts A, Gaizauskas R, Hepple M, et al. The CLEF corpus: semantic annotation of clinical text. AMIA Annu Symp Proc. 2007:625-9.
12. Chapman WW, Dowling JN, Hripcsak G. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. Int J Med Inform. 2008 Feb;77(2):107-13.
13. Denny JC, Spickard A, 3rd, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. J Am Med Inform Assoc. 2009 Nov-Dec;16(6):806-15.