Artificial Intelligence in Medicine (2004) **xxx**, xxx–xxx

ELSEVIER

Artificial Intelligence in Medicine

http://www.intl.elsevierhealth.com/journals/aiim

Classifying free-text triage chief complaints into syndromic categories with natural language processing

Wendy W. Chapman[a,*], Lee M. Christensen[b], Michael M. Wagner[a], Peter J. Haug[b], Oleg Ivanov[a], John N. Dowling[a], Robert T. Olszewski[a]

[a]*The RODS Laboratory, Center for Biomedical Informatics, University of Pittsburgh, Suite 8084, Forbes Tower, Pittsburgh, PA 15213, USA*
[b]*Department of Medical Informatics, LDS Hospital/University of Utah, 9th Avenue and C Street, Salt Lake City, UT 84143, USA*

Received 22 January 2004; received in revised form 26 March 2004; accepted 3 April 2004

KEYWORDS
Natural language processing;
Text classification;
Syndromic surveillance;
Chief complaints

Summary *Objective*: Develop and evaluate a natural language processing application for classifying chief complaints into syndromic categories for syndromic surveillance. *Introduction*: Much of the input data for artificial intelligence applications in the medical field are free-text patient medical records, including dictated medical reports and triage chief complaints. To be useful for automated systems, the free-text must be translated into encoded form. *Methods*: We implemented a biosurveillance detection system from Pennsylvania to monitor the 2002 Winter Olympic Games. Because input data was in free-text format, we used a natural language processing text classifier to automatically classify free-text triage chief complaints into syndromic categories used by the biosurveillance system. The classifier was trained on 4700 chief complaints from Pennsylvania. We evaluated the ability of the classifier to classify free-text chief complaints into syndromic categories with a test set of 800 chief complaints from Utah. *Results*: The classifier produced the following areas under the ROC curve: Constitutional = 0.95; Gastrointestinal = 0.97; Hemorrhagic = 0.99; Neurological = 0.96; Rash = 1.0; Respiratory = 0.99; Other = 0.96. Using information stored in the system's semantic model, we extracted from the Respiratory classifications lower respiratory complaints and lower respiratory complaints with fever with a precision of 0.97 and 0.96, respectively. *Conclusion*: Results suggest that a trainable natural language processing text classifier can accurately extract data from free-text chief complaints for biosurveillance.

1. Introduction

Much of the input data for artificial intelligence applications in the medical field are patient medical records, and a large portion of the medical record comprises free-text reports dictated by physicians, such as history and physical exams or radiology reports, that are unavailable for manipulation by computerized systems. A major goal of natural language processing (NLP) in the medical domain is to automatically extract and encode data stored in free-text patient records,

* Corresponding author. Tel.: +1 412 383 8140; fax: +1 412 383 8135.
E-mail address: chapman@cbmi.pitt.edu (W.W. Chapman).

0933-3657/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2004.04.001

essentially unlocking the information for use by computerized systems.

Over the past decade, research groups in the United States and Europe have been developing NLP systems to extract and encode information from medical reports [1—5] and have succeeded in restricted domains ranging from radiology reports to more complex discharge summaries [6—10].

In this project, we applied an NLP system designed for extraction of clinical information from dictated medical reports to the problem of classifying free-text triage chief complaints into syndromic categories useful for public health and bioterroristic outbreak detection. We adapted the NLP system to model chief complaints and evaluated its performance against classifications made by a physician with the objective of quantifying its performance at classifying chief complaints into syndromic categories.

2. Background

Early detection of bioterroristic or naturally-occurring disease outbreaks is crucial for saving lives [11,12]. Many diseases present themselves similarly in the early stages with non-specific symptoms that can be generalized into syndromic categories like respiratory or gastrointestinal. Syndromic surveillance is a type of early outbreak detection [13] that can be performed manually by public health officials or automatically by a computer.

The Real-time Outbreak and Disease Surveillance (RODS) System [14,15], initially deployed in Southwestern Pennsylvania in 1999, is a syndromic surveillance system that automatically monitors how frequently patients exhibit symptoms consistent with seven syndromes, including Botulinic, Constitutional, Gastrointestinal (GI), Hemorrhagic, Neurological, Rash, and Respiratory. Detection algorithms in RODS monitor complaints from patients presenting to emergency departments for unusual patterns of occurrence [16]. If the number of patients presenting to emergency departments with respiratory symptoms, for example, exceeds some threshold of expectation, RODS will alarm relevant medical and public health officials.

Input to RODS was initially ICD diagnostic codes [17] manually entered by a triage nurse from a patient's verbalized complaint. However, in February 2002, RODS expanded its reach beyond Southwestern Pennsylvania to monitor data from 30 urgent care facilities in Utah for the 2002 Winter Olympic Games [15,18,19]. Because manual coding of chief complaints into diagnostic codes is rare, we decided to use free-text triage chief complaints (TCCs), which are fairly ubiquitous across the nation, as input to RODS. A previous study showed that an NLP system can successfully encode free-text TCCs into diagnostic codes [20], therefore, we believed that the simpler problem of classifying TCCs into syndromic categories was likely to be successful.

Free-text triage chief complaints are the earliest clinical data available on most hospital information systems. A TCC is short phrase entered by a triage nurse describing the reason for a patient's visit to an emergency department. Some examples of common TCCs include ''cough,'' ''n/v/d,'' and ''luq abd pain.'' The purpose of a TCC is to describe a patient's condition in as short a space as possible, therefore, TCCs contain abbreviations and punctuation that can often confuse even experienced emergency room personnel. Researchers are investigating automated, knowledge-based methods for expanding TCC strings from abbreviated form into a more complete form [21,22].

2.1. The M+ system

The Medical Probabilistic Language Understanding System (M+) [23] is a robust chart-based syntactic parser with a Bayesian network (BN)-based semantic model for extracting information from narrative patient records. The semantic model can be trained in specific domains to adapt to new tasks. M+ has previously been applied to the domains of chest radiography [8] and brain CT scans.

M+ consists of the following components:

(1) A lexicon containing information largely derived from the UMLS Specialist Lexicon.
(2) A synonym component that maps common word and phrase variants to character strings matching states of word-level M+ BN nodes.
(3) Bayesian network-based semantic models for encoding words from the text and inferring concepts from combinations of words.
(4) A probabilistic spell checker combining an edit distance technique and ranking of the possible candidates based on probabilities generated by instantiations in the BNs.
(5) A semantic analyzer that instantiates BNs from phrases, passes information between BNs using virtual evidence, and generates semantic interpretations.
(6) A standard bottom-up chart parser with a context free grammar that seeks to create a deep parse. M+ BNs are instantiated during the syntactic parse. For example, as a word such as ''right'' is recognized by the parser, a word-level phrase object is created and a BN

instance containing the assignment *side* = *''right''* is attached to that phrase in the form of a predicate containing a token for that BN instance.

Each phrase recognized by the M+ parser is assigned a probability based on a weighted sum of the joint probabilities of its associated BN instances and adjusted for various syntactic and semantic constraint violations. Phrases are processed in order of probability, thus the parse involves a semantically-guided best-first search. Syntactic and semantic analyses in M+ are mutually constraining. On one hand, if a grammatically possible phrase is uninterpretable, i.e. if its subphrase interpretations cannot be unified, it is rejected. On the other hand, if the interpretation has a low probability, the phrase is less likely to appear in the final parse tree.

3. Methods

We adapted M+ to classify TCCs into syndromic categories and evaluated its performance against that of a human expert. This study was performed with IRB approval from relevant institutions in Utah and in Pennsylvania.

3.1. Definitions of syndromes

This paper describes a method for automatically encoding free-text TCCs into a standardized set of medical problems and then classifying them into the syndromes monitored by RODS. RODS's syndromic definitions were developed in 1999 and have only changed slightly since then, however, RODS syndromic definitions were initially comprised of lists of ICD-9 codes. Classifying free-text phrases into the syndromic categories raised questions about which chief complaints belong in which syndromes. As we developed the classification method described below, we generated heuristics for classification of specific complaints into syndromes. These heuristics were sometimes easily justifiable and other times quite arbitrary. Among other examples, decisions about appropriate classifications arose when a chief complaint represented a symptom that is not a very specific indicator of a public health outbreak or when a complaint could be due to more than one

1. **Gastrointestinal** includes pain or cramps anywhere in the abdomen, nausea vomiting, diarrhea and abdominal distension or swelling.
2. **Constitutional** is made up of non-localized, systemic problems including fever, chills, body aches, flu symptoms (viral syndrome), weakness, fatigue, anorexia, malaise, lethargy, sweating (diaphoresis), light headedness, faintness and fussiness. Shaking (not chills) is not constitutional but is other.
3. **Respiratory** includes the nose (coryza) and throat (pharyngitis), as well as the lungs. Examples of respiratory include congestion, sore throat, tonsillitis, sinusitis, cold symptoms, bronchitis, cough, shortness of breath, asthma, chronic obstructive pulmonary disease (COPD) and pneumonia. If both cold symptoms and flu symptoms are present, the syndrome is respiratory.
4. **Rash** includes any description of a rash, such as macular, papular, vesicular, petechial, purpuric or hives. Ulcerations are not normally considered a rash unless consistent with cutaneous anthrax (an ulcer with a black eschar).
5. **Hemorrhagic** is bleeding from any site, e.g., vomiting blood (hematemesis), nose bleed (epistaxis), hematuria, gastrointestinal bleeding (site unspecified), rectal bleeding and vaginal bleeding. Bleeding from a site for which we have a syndrome should be classified as hemorrhagic and as the relevant syndrome (e.g., Hematochesia is gastrointestinal and hemorrhagic; hemoptysis is respiratory and hemorrhagic).
6. **Botulinic** includes ocular abnormalities (diplopia, blurred vision, photophobia), difficulty speaking (dysphonia, dysarthria, slurred speech) and difficulty swallowing (dysphagia).
7. **Neurological** covers non-psychiatric complaints which relate to brain function. Included are headache, head pain, migraine, facial pain or numbness, seizure, tremor, convulsion, loss of consciousness, syncope, fainting, ataxia, confusion, disorientation, altered mental status, vertigo, concussion, meningitis, stiff neck, tingling and numbness. (Dizziness is constitutional and neurological.)
8. **Other** is a pain or process in a system or area RODS is not monitoring. For example, flank pain most likely arises from the genitourinary system, which RODS does not model, and would be considered other. Chest pain with no mention of the source of the pain is considered other (e.g., chest pain (other) versus pleuritic chest pain (respiratory)). Earache or ear pain is other. Trauma is other.

Figure 1 Syndromic definitions used by RODS and examples of chief complaints that should be classified into the syndromes.

syndrome. Fig. 1 contains the final definitions of syndromes we developed for this task.

Since this project, we have begun to evaluate the quality of the syndromic definitions by measuring how well TCCs classified into the syndromes can detect patients who actually have the syndromes. Using physician review of emergency department reports as a reference standard, we can detect patients with acute lower respiratory syndrome by manually classifying their TCCs into our respiratory syndrome with a sensitivity of 0.80 and a specificity of 0.94 (unpublished results) and can detect patients with an acute, infectious gastrointestinal illness with a sensitivity of 0.63 and a specificity of 0.94 [24]. We are currently evaluating the quality of our syndromic definitions at classifying patients into all seven syndromes monitored by RODS. Moreover, we have shown that the syndromic categories are sensitive and specific indicators of outbreaks of respiratory and gastrointestinal illness in children [25].

3.2. Adapting M+ to the current project

Because TCC strings often contain abbreviations, common abbreviations were included in a manually created synonym list. For instance, the abbreviation ''t.i.a.'' is expanded to ''transient ischemic attack'' which maps to one or more BN states. Spell checking was disabled in this project for the reason that the most common misspellings in TCCs are due to truncations of the final word or to abbreviations which do not lend themselves well to a meaningful edit distance measure of word similarity. Therefore, common misspellings, truncations, and abbreviations were also handled through the synonym list.

We used a single BN to model the semantic information described in TCCs, illustrated in Fig. 2 and described in Section 3.3.

M+ was originally designed to analyze narrative and descriptive medical reports, such as radiology dictations. Since the texts in the current study are grammatically and lexically abbreviated, a simplified grammar containing rules for conjunctions and simple noun, adjective, and prepositional phrases was used. We also added a probabilistic algorithm for identifying noun and adjectival phrases within contiguous sets of words, without the use of rules.

3.3. Training M+

M+'s semantic component is based on the premise that the number of ideas or concepts expressed by language is fewer than the actual words used to express the concepts. Bayesian network-based semantics [23] is analogous to Bayesian network-based diagnosis in which the top parent node is a joint probability distribution over all possible diseases and the child nodes are a patient's symptoms or findings. In the diagnostic domain, the same disease can manifest itself with various combinations of different findings. Similar to medical diagnosis, the concepts being expressed in medical texts are manifested using various combinations of different words. In a Bayesian network-based semantic model, the parent nodes represent the underlying concept expressed by the author of the text and the child nodes represent the words used to express those concepts. For example, a Bayesian network

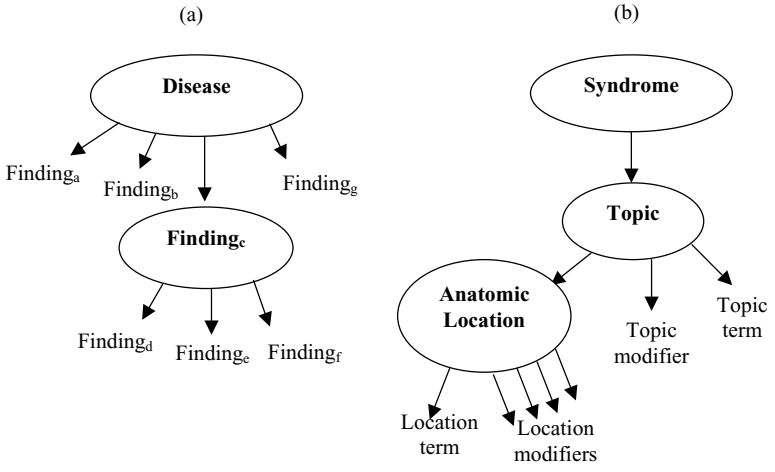

Figure 2 Similar to a diagnostic Bayesian network (a), the semantic component for M+ (b) uses a Bayesian network that stores words from the input text (leaf nodes) and concepts that can be inferred from the words. The top node is a joint probability distribution over all possible syndromes. The structure of the network was created by the authors, but the parameters were learned from training examples.

for chest radiology reports would classify an ''ill-defined density,'' a ''hazy opacity,'' and a ''patchy opacification'' as the same underlying concept of ''pneumonic infiltrate.''

We created a Bayesian network to represent the relationship of words and concepts expressed in TCCs, shown in Fig. 2. We constructed the network by hand, but the network's parameters were learned with training data. Two of the authors (WWC and OI) trained M+ with 4700 randomly selected TCCs from emergency department admissions in one hospital in Pittsburgh, PA.

A Bayesian network-based semantic model can be a powerful model for NLP, but providing training examples for the domain of TCCs involves more than assigning the correct syndromic classification to an individual string of text. Instantiating the training example shown in Fig. 3 requires not only specifying the syndromic classification of the string but also slotting relevant words from the TCC in the correct nodes and creating appropriate concepts that represent combinations of words. A web-based training tool created by one of the authors (LMC) was used to train M+ for chief complaint instantiation. The training tool used the current training set to guess the correct instantiation for a new training case. Users of the tool could accept or change the tool's instantiation and could select from concepts already used in other training cases. Suggestions from the web training tool sped up the training process, facilitated application of consistent concepts to the training cases, and helped the trainers be consistent with each other and with themselves.

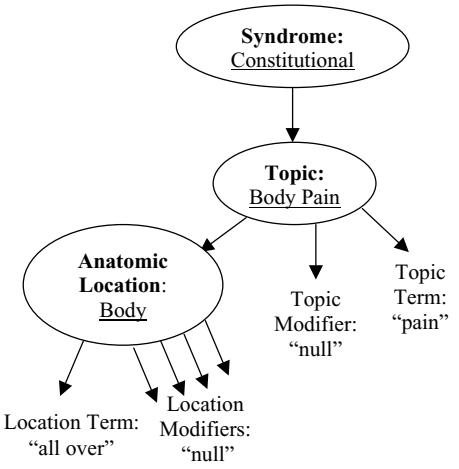

Figure 3 An instantiation of a training example for the string ''feeling pain all over.'' The authors training the system specified the Location Term ''all over'' to represent the Anatomic Location Concept of Body. An Anatomic Location Concept of Body combined with the Topic Term ''pain'' was specified as the Topic Body Pain. Body Pain was specified to belong to the Constitutional category.

3.4. Output of M+

Once the Bayesian network was trained on the training cases, M+ was able to classify unseen TCCs into syndromic categories. For this task, M+'s output is a joint probability distribution over all possible values for the Syndrome node. Possible values for the Syndrome node include the seven syndromes monitored by RODS and the value Other. TCCs classified as Other are those complaints which are not of concern in outbreak detection, including trauma and complaints involving the musculoskeletal system and genitourinary tract.

Sometimes a single TCC describes more than one medical problem, as in ''cough and vomiting.'' Because M+ performs syntactic and semantic analyses, M+ is potentially able to detect multiple problems from one phrase. In the example above, M+'s output would include multiple probability distributions, one for each problem detected (e.g. a probability distribution for ''cough'' and a separate distribution for ''vomiting''). Therefore, M+ may classify one TCC into more than one syndrome.

In production, a probabilistic threshold can be applied to M+'s output to determine the most probable syndrome(s) for the TCC. The threshold can be chosen to reflect the user's preference of high sensitivity or high specificity. In this study, we evaluated M+'s classification performance with the area under the ROC curve (AUC), which measures performance at all probabilistic thresholds [26].

3.5. Evaluation

We evaluated M+'s ability to accurately classify free-text TCCs into eight syndromic categories by comparing M+'s classifications against those of a gold standard physician. The gold standard physician, who is board-certified in internal medicine and infectious diseases, also acted as a medical consultant in defining the syndromic definitions. We evaluated the validity of the gold standard responses by comparing his classifications of the 800 TCCs against one of the author-trainers (OI) for agreement to ensure that the gold standard physician and the system were performing the same task, an assumption necessary when using a human expert as the gold standard [27].

The test set contained the first 1000 TCCs received from urgent care facilities in Utah (beginning January 29, 2002). The first 200 TCCs were used to train the gold standard physician on the

task. We trained the physician by describing the final case definitions shown in Fig. 1 and giving him feedback on his classifications of the 200 training cases.

M+ and the gold standard physician independently classified the remaining 800 TCCs. Predictive performance of M+ was evaluated by calculating the AUC using trapezoidal integration. We also report sensitivity, specificity, and positive predictive value (PPV) for all classification thresholds. Because a single TCC could be classified into multiple syndromes, we calculated the AUC for every syndrome individually. In this way if M+ classified ''cough and vomiting'' as Respiratory, but the gold standard physician classified the string as Respiratory and GI, the TCC would contribute a true positive count to the AUC calculation for Respiratory and a false negative count to the AUC for GI.

When applying probabilistic thresholds to M+'s output in order to plot points on the ROC curve, we incorporated four guidelines created by two physician authors (OI and JND) for assigning multiple syndromic classifications to a single TCC. The first three guidelines are based on medical knowledge about co-occurring conditions and the fourth is based on the fact that RODS does not monitor medical problems classified as Other.

(1) If part of the TCC describes a motor vehicle accident (mva), the entire TCC is explained away by the mva and should be classified as Other (e.g. ''headache from mva'' would only be classified as Other even though ''headache'' is classified as Neurological).
(2) A problem classified as Botulinic is explained away by any other co-occurring syndromic classification (e.g. ''sore throat/difficulty swallowing'' would only be classified as Respiratory even though ''difficulty swallowing'' is classified as Botulinic).
(3) A fever is explained away by any other co-occurring syndromic classification (e.g. ''cough, fever'' would only be classified as Respiratory even though ''fever'' is classified as Constitutional).
(4) A classification of Other cannot be combined with another syndromic classification (e.g. ''diarrhea and broken arm'' would only be classified as GI).

Because M+ stores semantic information about the TCCs in the Bayesian network structure, we can extract more specific information from the string than the syndromic classification. From the words in the string M+ infers the topic of the complaint, so we can divide patients classified by a syndrome into subsets of topics (i.e. complaints) that may be helpful as a first step to investige a possible outbreak. For example, the first question one may ask when investigating a possible respiratory outbreak is whether the cases classified by M+ as Respiratory presented with upper or lower respiratory symptoms, because lower respiratory symptoms are more likely to be caused by an agent of concern to a public health investigation. Patients with a lower respiratory complaint and a fever would be even more concerning.

We broke down the test set TCCs classified by M+ as Respiratory into Lower Respiratory and Lower Respiratory with Fever subcategories based on the Topic concepts output by M+. Looking only at M+'s output and not at the original string, all Respiratory cases with a Topic concept involving the lower respiratory tract (e.g., Cough, Pneumonia, Dyspnea, etc.) were labeled as Lower Respiratory. From the Lower Respiratory subcategory, we reviewed all cases for which M+ generated multiple complaints. If one of the complaints had a Topic concept of Fever, we labeled the case as Lower Respiratory with Fever. The gold standard physician then read all the TCC strings from the cases classified by M+ as Respiratory. Based on the string itself, he labeled the cases as either Lower Respiratory or Lower Respiratory with Fever. We compared classifications made from M+'s output to classifications made by the gold standard physician from the TCC string to calculate the precision (positive predictive value) of classifying the two subcategories from M+'s Topic concept.

4. Results

Fig. 4 shows the distribution of test set classifications made by the gold standard physician. None of the test TCCs were classified as Botulinic. More than 50% of the test set was classified as Other. We measured the Cronbach α reliability coefficients

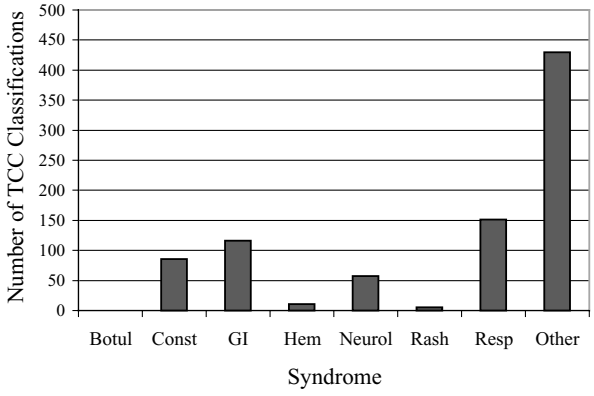

Figure 4 Distribution of TCC classifications by the gold standard physician.

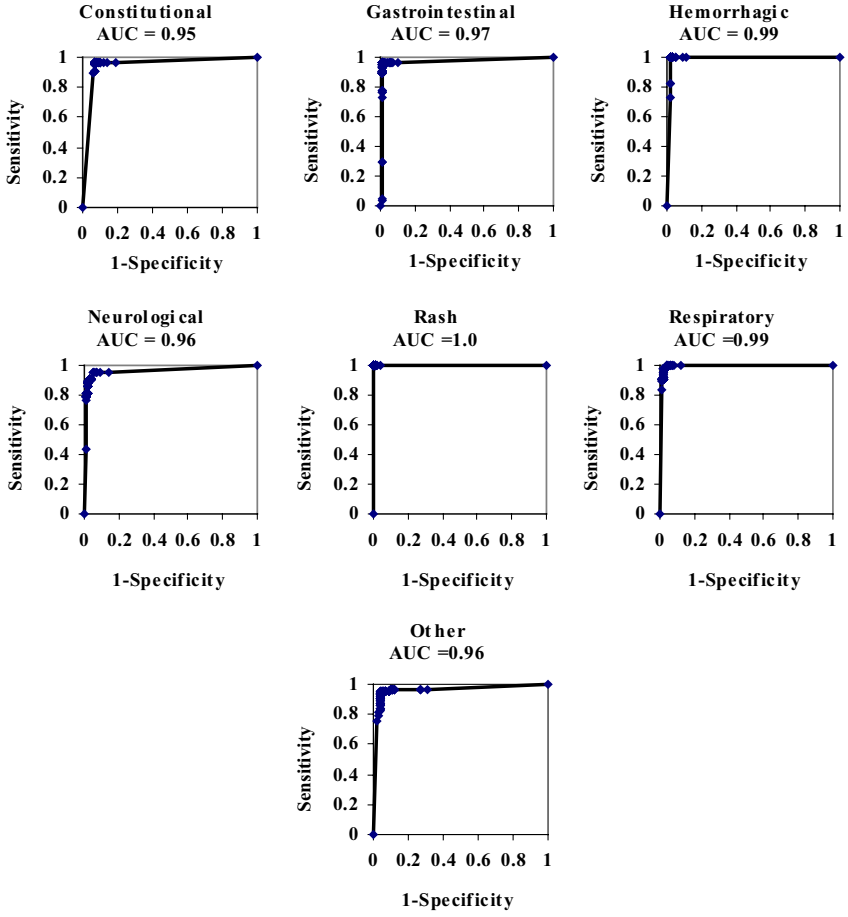

Figure 5 Receiver Operator Characteristic (ROC) curves generated from comparing M+'s classifications to those of the gold standard physician for the seven syndromic categories occurring in the test set. The area under the curve (AUC) is reported for every syndrome.

[28] for agreement between the gold standard physician and an author trainer. Most of the syndromic categories received high coefficients: GI = 0.96; Respiratory = 0.94; Other = 0.94; Constitutional = 0.93; Neurological = 0.86. Agreement was fair between the gold standard physician and the training physician on Rash (0.71) and Hemorrhagic (0.77).

Fig. 5 shows the ROC curves calculated from comparing M+'s classifications against those of the gold standard physician for every syndrome except Botulism, which did not occur in the test set. AUC's ranged between 0.95 (Constitutional) and 1.0 (Rash). Because sensitivity is probably the most important measure in an outbreak detection system that is meant to screen for syndromic presentations, we show in Table 1, the specificity and PPV for probabilistic thresholds that yield M+'s highest classification sensitivity.

Table 1 M+'s predictive performance for probabilistic thresholds yielding highest sensitivity

Syndrome	Sensitivity	Specificity	PPV
Constitutional	0.97	0.93	0.97 (83/86)
GI	0.96	0.99	0.96 (111/116)
Hemorrhagic	1.0	0.98	0.91 (10/11)
Neurological	0.95	0.95	0.95 (54/57)
Rash	1.0	0.99	0.8 (4/5)
Respiratory	1.0	0.96	0.87 (131/151)
Other	0.97	0.88	0.92 (396/430)

Fig. 6 illustrates the breakdown of the 150 TCCs classified by M+ as Respiratory into Lower Respiratory (98/150) and Lower Respiratory with Fever (26/98). M+ extracted Lower Respiratory complaints with a precision of 0.97 (95/98) and Lower Respiratory with Fever with a precision of 0.96 (25/26).

5. Discussion

As shown in Fig. 5, M+ performed well at classifying TCCs into the seven syndromes represented in the test set with AUC's ranging from 0.95 to 1.0. Probabilistic thresholds yielding the highest sensitivity generated sensitivity rates between 0.95 and 1.0

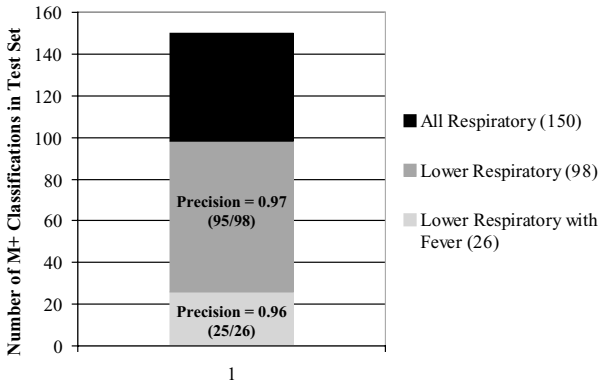

Figure 6 Breakdown of test cases classified by M+ as Respiratory. Of the 150 respiratory classifications, 98 were labeled as Lower Respiratory with a precision of 0.97 (95/98). Of the 98 Lower Respiratory classifications, 26 were labeled as Lower Respiratory with Fever with a precision of 0.96 (25/26).

with corresponding specificity rates of at least 0.88. These results suggest that M+ can accurately classify free-text chief complaint strings into syndromic categories.

An error analysis on M+'s performance revealed a few common sources of mistakes. First, because M+ is trained on a finite set of training examples, some instances in the test set, like ''croup'' and ''RSV,'' were previously unseen by M+. Fortunately, M+ is a trainable system whose performance will potentially improve with more training. We were surprised that a classifier trained completely on TCCs from one adult hospital in Pittsburgh performed as well as it did on TCCs from 30 urgent care facilities, both adults' and children's, in Utah.

Second, several of M+'s mistakes occurred because we are classifying TCCs into a non-exhaustive set of syndromes that classifies medical problems irrelevant to biosurveillance in the category Other. Irrelevant and relevant TCCs have substantial overlap in their lexical content. For example, M+ was not trained on the term ''respiratory problem'' but was trained on ''eye problem,'' which was classified as Other. In classifying the TCC ''respiratory problem,'' M+ assigned Other a higher probability than Respiratory. We are currently evaluating an active learning technique that uses the semantic knowledge contained in the training set to select new training cases with the highest expected value to the classifier.

Very few mistakes in the test set were due to misspellings, however, there is no way to know in advance all misspellings that may occur [29,30]. Future work includes incorporating the existing spell checker using words and phrases in M+'s training set as the dictionary and adding to the synonym list abbreviations and spelling variants from the UMLS Metathesaurus [31]. Neither of these approaches eliminates the need for a synonym list for translating uncommon abbreviations frequently seen in TCCs. For example, ''appy'' for ''appendicitis'' and ''tib fib'' for ''tibia with fibula'' frequently occur in TCCs but could not be translated correctly with a spell checker or from the Metathesaurus.

We measured M+'s performance by comparing its classifications against classifications made by a gold standard physician. Using expert physician judgment as a reference standard is a common method in evaluating the performance of medical classification system [27], because AI systems in medicine are often designed to imitate physician performance. Comparing an AI system against an expert-generated reference standard assumes that the expert is performing the same task as the system [27]. Some medical classification problems, such as clinical diagnosis, are well defined and understood outside of any research project, but to our knowledge the task of classifying TCCs into syndromes for the purpose of measuring performance of an automated classification system has not been done before. Moreover, even within the syndromic surveillance community, syndromic definitions of the same syndrome (e.g. Respiratory syndrome) still vary substantially [32].

Because syndromic definitions are so dependant on the surveillance system, creating reliable gold standard classifications for a syndromic classifier is difficult. Evaluations of NLP systems using physician judgment as a reference standard typically employ multiple physician judgments [27]. It can be argued, however, that acting as a gold standard for this task requires project-specific expertise about syndromic definitions. To ensure that the gold standard physician was performing the same task as M+, we compared his classifications against classifications made by the training physician, and their agreement was quite high on five of the seven syndromes (Cronbach α reliability coefficients from 0.86 to 0.96).

We believe that our gold standard of one physician is reasonable for this task. We are currently examining the possibility of training people outside our research group either to supplement the one-person gold standard or to help decrease the noise that undoubtedly occurs when only one person classifies TCCs into syndromic categories.

5.1. Using M+ for classification of TCCs

We adapted M+ to the domain of biosurveillance from TCCs in order to both classify the TCCs into the

syndromic categories monitored by RODS and encode TCCs into a standardized subset of medical problems. The goal of text classification is to automatically classify a set of documents into one of a discrete set of possible categories [33]. A variety of text classification algorithms [34–36] have successfully classified documents such as Medline abstracts [37] and web pages using algorithms like support vector machines, k-nearest-neighbor, and Naïve Bayes'. Most approaches to text classification involve minimal, if any, syntactic processing of the text.

Other syndromic surveillance systems are also beginning to use free-text TCCs and are typically using keyword-matching methods for classifying the complaints into syndromes. RODS has also implemented a simpler, naïve Bayesian TCC classifier called CoCo [38]. CoCo was trained on 10,000 TCCs from Utah. In an evaluation using 10-fold cross validation, CoCo classified the TCCs into the same syndromes described above and received AUC scores between 0.78 and 0.97. The question still remains whether the overhead of syntactic parsing and Bayesian semantic analysis is necessary for classifying TCCs into syndromes—the loss in performance of a naïve Bayesian classifier may be worth the simpler and more portable architecture. Moreover, more advanced classifiers may outperform naïve Bayesian classification and approach M+'s accuracy without syntactic parsing. A shortcoming of this paper is a lack of a direct comparison to a standard text classification system. The explanation for this shortcoming is our lack of foresight. We had 6 weeks to develop and implement a classification system for the Olympics, and we did not consistently maintain a link between the training cases and the text of the TCCs represented by the training cases. For this reason, we were not able to train another text classification system on the same training data for direct comparison.

M+'s more complex structure has several advantages over simple text classification techniques. First, the syntactic and semantic analyses help analyze coordinations in the TCCs (e.g. ''rt shoulder/arm pain'' or ''neck pain numb rt arm''), so that multiple complaints can result in multiple syndromic classifications, if appropriate. Second, in addition to generating a syndromic classification, M+ can also encode the specific medical problem described in the chief complaint so that investigators of an outbreak could query the TCCs for further information. We illustrated this feature by using M+'s Topic concepts to break down Respiratory complaints into subcategories of Lower Respiratory and Lower Respiratory with Fever. As shown in Fig. 6, M+ was able to classify cases into these subcategories with high precision.

6. Conclusion

Using free-text chief complaints to classify patients into syndromic categories is a new approach to syndromic surveillance that can potentially provide real-time, early clinical data from the entire nation. To be useful, however, the textual phrases must first be classified into syndromic categories. Results of this study suggest that a trainable natural language processing system can successfully classify triage chief complaints into syndromes. Moreover, the semantic model of our system can provide detailed information that can be easily queried to aid investigators in an initial attempt to understand the nature of the syndromic cases, making natural language processing a potentially valuable tool in national outbreak detection.

Acknowledgements

We thank Zhongwei Lu for programming help. This work was in part supported by NLM training grant T15 LM07059 and CDC grant UPO/CCU 318753-02.

References

[1] Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1994;1:142–60.
[2] Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999;74:890–5.
[3] Friedman C. A broad-coverage natural language processing system. In: Proceedings of the AMIA Symposium; 2000. p. 270–4.
[4] Taira RK, Soderland SG, Jakobovits RM. Automatic structuring of radiology free-text reports. Radiographics 2001;21:237–45.
[5] Baud RH, Lovis C, Ruch P, Rassinoux AM. A toolset for medical text processing. Stud Health Technol Inform 2000;77:456–61.
[6] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 1995;122:681–8.
[7] Hahn U, Schnattinger K, Romacker M. Automatic knowledge acquisition from medical texts. In: Proceedings of AMIA Annual Fall Symposium; 1996. p. 383–7.
[8] Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000;7:593–604.
[9] Friedman C, Knirsch C, Shagina L, Hripcsak G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. In: Proceedings of the AMIA Symposium; 1999. p. 256–60.
[10] Hripcsak G, Austin JH, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 2002;224:157–63.

[11] Wagner MM, Tsui FC, Espino JU, Dato VM, Sittig DF, Caruana RA et al. The emerging science of very early detection of disease outbreaks. J Public Health Manag Pract 2001;7:51–9.
[12] Hashimoto S, Murakami Y, Taniguchi K, Nagai M. Detection of epidemics in their early stage through infectious disease surveillance. Int J Epidemiol 2000;29:905–10.
[13] Paviln JA. Electronic surveillance system for the early notification of community-based epidemics (ESSENCE). In: Conference and Workshop on Syndromic and Other Surveillance Methods for Emerging Infections Including Bioterrorism; 2000.
[14] Tsui FC, Espino JU, Dato VM, Gesteland PH, Hutman J, Wagner MM. Technical description of RODS: a real-time public health surveillance system. J Am Med Inform Assoc 2003;10:399–408.
[15] Tsui FC, Espino JU, Wagner MM, Gesteland P, Ivanov O, Olszewski R, et al. Data, network, and application: technical description of the Utah RODS winter olympic biosurveillance system. In: Proceedings of the AMIA Symposium; 2002. p. 815–9.
[16] Wong W, Moore AW, Cooper G, Wagner M. Rule-based anomaly pattern detection for detecting disease outbreaks. In: Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-02); 2002.
[17] Tsui FC, Wagner MM, Dato V, Chang CC. Value of ICD-9 coded chief complaints for detection of epidemics. In: Proceedings of the AMIA Symposium; 2001. p. 711–5.
[18] Gesteland PH, Wagner MM, Chapman WW, Espino JU, Tsui F, Gardner RM, et al. Rapid deployment of an electronic disease surveillance system in the state of utah for the 2002 olympic winter games. In: Proceedings of the AMIA Symposium; 2002. p. 285–9.
[19] Gesteland PH, Gardner RM, Tsui FC, Espino JU, Rolfs RT, James BC. Automated syndromic surveillance for the 2002 Winter Olympics. J Am Med Inform Assoc 2003;10:547–54.
[20] Haug PJ, Christensen L, Gundersen M, Clemons B, Koehler S, Bauer K. A natural language parsing system for encoding admitting diagnoses. In: Proceedings of the AMIA Annual Fall Symposium; 1997. p. 814–8.
[21] Travers DA, Haas SW. Using nurse's natural language entries to build a concept-oriented terminology for patient's chief complaints in the emergency department. J Biomed Inform 2003;36:260–70.
[22] Travers DA, Waller A, Haas SW, Lober WB, Beard C. Emergency department data for bioterrorism surveillance: electronic data availability, timeliness, sources and standards. In: Proceedings of the AMIA Symposium; 2003. p. 664–8.
[23] Christensen L, Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding system. In: Proceedings and Workshop on Natural Language Processing in the Biomedical Domain; 2002. p. 29–36.
[24] Ivanov O, Wagner MM, Chapman WW, Olszewski RT. Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. In: Proceedings of the AMIA Symposium; 2002. p. 345–9.
[25] Ivanov O, Gesteland P, Hogan W, Mundorff MB, Wagner MM. Detection of pediatric respiratory and gastrointestinal outbreaks from free-text chief complaints. In: Proceedings of the AMIA Annual Fall Symposium; 2003. p. 318–22.
[26] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
[27] Hripcsak G, Wilcox A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc 2002;9:1–15.
[28] Friedman C, Wyatt JC. Evaluation methods for medical informatics. New York: Springer-Verlag; 1997.
[29] Ruch P, Baud R, Geissbuhler A, Rassinoux AM. Comparing general and medical texts for information retrieval based on natural language processing: an inquiry into lexical disambiguation. Medinfo 2001;10:261–5.
[30] Ruch P, Baud R, Geissbuhler A. Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records. Int J Med Inf 2002;67:75–83.
[31] McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med 1995;34:193–201.
[32] Graham J, Buckeridge D, Choy M, Musen M. Conceptual heterogeneity complicates automated syndromic surveillance for bioterrorism. In: Proceedings of the AMIA Annual Fall Symposium; 2002. p. 1030.
[33] Mitchell TM. Machine learning. Boston, MA: The McGraw-Hill Companies Inc., 1997.
[34] Yang YM. An evaluation of statistical approaches to text categorization. J Inform Retrieval 1999;1:67–88.
[35] Yang YM, Liu X. A re-examination of text categorization methods. In: Proceedings of the ACM SIGIR; 1999. p. 42–9.
[36] Lam W, Ho CY. Using a generalized instance set for automatic text categorization. In: Proceedings of the ACM SIGIR; 1998. p. 81–9.
[37] Yang YM. An evaluation of statistical approaches to MEDLINE indexing. In: Proceedings of the AMIA Symposium; 1996. p. 358–62.
[38] Olszewski RT. Bayesian classification of triage diagnoses for the early detection of epidemics. In: Proceedings of the FLAIRS Conference; 2003. p. 412–6.