ELSEVIER

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Multilingual chief complaint classification for syndromic surveillance: An experiment with Chinese chief complaints

Hsin-Min Lu[a,*], Hsinchun Chen[a], Daniel Zeng[a,d], Chwan-Chuen King[b], Fuh-Yuan Shih[c], Tsung-Shu Wu[b], Jin-Yi Hsiao[b]

[a] Management Information Systems Department, Eller College of Management, University of Arizona, 1130 East Helen Street, McClelland Hall 430, Tucson, Arizona 85721, USA
[b] Graduate Institute of Epidemiology, National Taiwan University, Taipei, Taiwan
[c] Department of Emergency Medicine, National Taiwan University Hospital, No. 7, Chung-Shan South Road, Taipei 100, Taiwan
[d] Institute of Automation, Chinese Academy of Sciences, Zhongguancun East Road #95, Beijing, China

ARTICLE INFO

Article history:
Received 3 January 2008
Received in revised form
18 August 2008
Accepted 19 August 2008

Keywords:
Communicable disease control
Medical records
Multilingual chief complaint classification
Syndromic surveillance
Statistical pattern extraction

ABSTRACT

Purpose: Syndromic surveillance is aimed at early detection of disease outbreaks. An important data source for syndromic surveillance is free-text chief complaints (CCs), which may be recorded in different languages. For automated syndromic surveillance, CCs must be classified into predefined syndromic categories to facilitate subsequent data aggregation and analysis. Despite the fact that syndromic surveillance is largely an international effort, existing CC classification systems do not provide adequate support for processing CCs recorded in non-English languages. This paper reports a multilingual CC classification effort, focusing on CCs recorded in Chinese.

Methods: We propose a novel Chinese CC classification system leveraging a Chinese-English translation module and an existing English CC classification approach. A set of 470 Chinese key phrases was extracted from about one million Chinese CC records using statistical methods. Based on the extracted key phrases, the system translates Chinese text into English and classifies the translated CCs to syndromic categories using an existing English CC classification system.

Results: Compared to alternative approaches using a bilingual dictionary and a general-purpose machine translation system, our approach performs significantly better in terms of positive predictive value (PPV or precision), sensitivity (recall), specificity, and F measure (the harmonic mean of PPV and sensitivity), based on a computational experiment using real-world CC records.

Conclusions: Our design provides satisfactory performance in classifying Chinese CCs into syndromic categories for public health surveillance. The overall design of our system also points out a potentially fruitful direction for multilingual CC systems that need to handle languages beyond English and Chinese.

[*] *Corresponding author*. Tel.: +1 520 621 2165.
E-mail addresses: hmlu@email.arizona.edu (H.-M. Lu), hchen@eller.arizona.edu (H. Chen), zeng@eller.arizona.edu (D. Zeng), cc_king99@hotmail.com (C.-C. King), wcsg@msn.com (T.-S. Wu), pingeshity@yahoo.com.tw (J.-Y. Hsiao).

1386-5056/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2008.08.004

1. Introduction

Modern transportation shortens the time needed for a person to travel from one side of the globe to the other. At the same time, it also shortens the time needed for a disease to spread. A case in point is the severe acute respiratory syndrome (SARS) episode which started in the Guangdong Province, China in November, 2002 and spread to Toronto, Vancouver, Ulaan Bator, Manila, Singapore, Hanoi, and Taiwan by March, 2003. The disease was finally brought under control and the whole episode ended in July, 2003. There were a total of 8096 known cases, and about 35% were outside mainland China (cf. http://www.who.int/csr/sars/en/).

The SARS experience indicates that an effective plan for infectious disease detection and prevention, in which syndromic surveillance may play an important role, should be considered on a global scale [1,2]. However, only a few countries have adopted formal syndromic surveillance systems. The U.S. public health system has significant experience in developing and adopting syndromic surveillance systems. However, leveraging such experience in international contexts is proven to be difficult. Multilingual data present a major barrier, as different languages are used by medical and public health practitioners in different parts of the world. This is particularly true for a major data source used by many syndromic surveillance systems: emergency department (ED) triage free-text chief complaints (CCs).

ED triage free-text CCs are short free-text phrases entered by triage practitioners describing reasons for patients' ED visits. ED CCs are a popular data source because of their timeliness and availability [3–6]. However, medical practitioners in other countries do not always use English when recording patients' CCs [7]. As a result, existing CC classification systems designed for English CCs cannot be directly applied in these countries as an important component of the overall syndromic surveillance strategy.

For automatic syndromic surveillance, free-text CC records need to be classified into predefined syndromic categories. This paper reports a study examining the importance of Chinese CCs as a data source for syndromic surveillance and aims to develop a Chinese CC syndromic classification approach. This research was motivated to answer the following research questions:

(a) How useful Chinese CCs are for syndromic surveillance and
(b) Whether an effective cross-lingual approach can be developed leveraging existing English CC classification methods.

CCs from EDs in Taiwan were collected and analyzed in our research. Medical practitioners in Taiwan are trained to record CCs in English. However, it is a common practice to record CCs in both Chinese and English. Furthermore, some hospitals record CCs only in Chinese. We systematically investigated the role and validity of Chinese CCs in the syndromic surveillance context. We then developed a system to classify Chinese CCs based on an automated mechanism to map Chinese CCs to English CCs.

The remainder of this paper is organized as follows. Section 2 provides the background for existing CC classification and cross-lingual information retrieval methods. The next section presents research opportunities and objectives of our research. Section 4 describes our findings regarding the importance of Chinese CCs. Sections 5 and 6 discuss system design of the Chinese CC classification system and experiments to study system performance. Section 7 concludes our discussion.

2. Research background

This section reviews existing CC classification research for both English and non-English CCs. Cross-lingual information retrieval and Chinese key phrase extraction and text segmentation are also reviewed as it provides technical foundation for this research.

2.1. *English chief complaint classification methods*

There are three main approaches for automated CC syndrome classification: supervised learning, rule-based classification, and ontology-enhanced classification. The supervised learning methods require CC records to be labeled with syndromes before being used for model training. Naive Bayesian [8–10] and Bayesian network [4] models are two examples of the supervised learning methods studied. One prerequisite of supervised learning methods is collecting a sufficient amount of training records, which is usually costly and time-consuming. Another major disadvantage of supervised learning methods is the lack of flexibility. New syndromic definitions may be required by public health practitioners as new events may indicate new surveillance focuses. However, it is often difficult to produce new training data for new syndromic definitions.

Rule-based classification methods do not require labeled training data. Such methods typically have two stages. In the first stage, CC records are cleaned up and transformed to an intermediate representation called "symptom groups" by either a symptom grouping table (SGT) lookup or keyword matching. In the second stage, a set of rules is used to map the intermediate symptom groups to final syndromic categories. For instance, the EARS system (http://www.bt.cdc.gov/surveillance/ears/) uses 42 rules for such mappings.

A major advantage of rule-based classification methods is their simplicity. The syndrome classification rules and intermediate SGTs can be constructed using a top-down approach. The "white box" nature of these methods makes system maintenance and fine-tuning easy for system designers and users. In addition, these methods are flexible. Adding new syndromic categories or changing syndromic definitions can be achieved relatively easily by switching the inference rules. The SGTs can typically be shared across hospitals.

A major problem with rule-based classification methods is that they cannot handle symptoms that are not included in the SGTs. For example, a rule-based system may have a SGT containing the symptoms "abdominal pain" and "stomach ache." This system, however, will not be able to handle "epi-