Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Risk factor detection for heart disease by applying text analytics in electronic medical records

Manabu Torii *, Jung-wei Fan, Wei-li Yang, Theodore Lee, Matthew T. Wiley, Daniel S. Zisook, Yang Huang

Medical Informatics, Kaiser Permanente Southern California, 11975 El Camino Real, Suite 105, San Diego, CA, United States

ARTICLE INFO

Article history:
Received 14 February 2015
Revised 6 August 2015
Accepted 7 August 2015
Available online xxxx

Keywords:
Medical records
Risk assessment
Natural language processing
Text classification

ABSTRACT

In the United States, about 600,000 people die of heart disease every year. The annual cost of care services, medications, and lost productivity reportedly exceeds 108.9 billion dollars. Effective disease risk assessment is critical to prevention, care, and treatment planning. Recent advancements in text analytics have opened up new possibilities of using the rich information in electronic medical records (EMRs) to identify relevant risk factors. The 2014 i2b2/UTHealth Challenge brought together researchers and practitioners of clinical natural language processing (NLP) to tackle the identification of heart disease risk factors reported in EMRs. We participated in this track and developed an NLP system by leveraging existing tools and resources, both public and proprietary. Our system was a hybrid of several machine-learning and rule-based components. The system achieved an overall F1 score of 0.9185, with a recall of 0.9409 and a precision of 0.8972.

© 2015 Elsevier Inc. All rights reserved.

1. Background

In the United States, heart disease is the leading cause of death, accounting for over 600,000 deaths per year [1]. The American Heart Association reports that the annual total cost of care services, medications, and lost productivity exceeds 108.9 billion dollars [2]. The 2014 i2b2/UTHealth Challenge brought together researchers and practitioners of clinical natural language processing (NLP) to tackle problems of common interest, which included the identification of heart disease risk factors reported in electronic medical records (EMRs), a task that will support prevention, care, and treatment planning of the disease. We participated in a track focusing on this task, Track 2 of the 2014 i2b2/UTHealth Challenge.

The goal of the Track-2 task was to annotate diagnoses, risk factors, and associated medications at the record (document) level. The challenge organizer provided a training corpus with gold annotations at the record level, but also made available raw evidence annotations at the phrase level for participants' system development. The task concerns several clinical NLP topics including disease concept identification, medication detection, and smoking status classification. Each of these topics may require supporting tasks, such as assertion detection, section detection, and temporal information detection, as well as basic NLP tasks, such as part-of-speech tagging. Many of these tasks have been studied over the years [3–14]. Track 2 of the 2014 i2b2/UTHealth Challenge provided a valuable opportunity to determine the generalizability of past research.

Because significant overlap exists between the 2014 Challenge and prior i2b2 Challenges, we reviewed successful prior efforts, and discovered a common technique, which was termed "hot-spot identification" by Cohen [12]. In this technique, a small amount of discriminative words are identified to classify a document. The hot-spot phrases may be identified via hand-coded rules or sequence labeling techniques, such as Conditional Random Field (CRF) [8]. Hot-spot-based techniques have demonstrated repeated success by multiple teams during the 2006 and 2011 i2b2 Challenges.

In the 2006 i2b2 NLP Challenge, hot-spot phrases were leveraged to classify patients' smoking status [9] (*non-smoker*, *current smoker*, *past smoker*, *smoker*, or *unknown*). Among the best performing systems in the challenge were those developed by Aramaki et al. [10], Clark et al. [11], and Cohen [12], which achieved micro-averaged *F*-measures of 0.88, 0.90, and 0.89, respectively. All of these top performers used hot-spot-based techniques. Aramaki et al. tackled this task in two steps. In the first step, a single sentence reporting the patient's smoking status was selected from a medical record. This selection was based on the

* Corresponding author at: Medical Informatics, Kaiser Permanente Southern California, 11975 El Camino Real, Suite 105, San Diego, CA 92130, United States. Tel.: +1 (858) 523 6409; fax: +1 (858) 523 6423.
E-mail addresses: manabu.torii@kp.org, manabu.torii@gmail.com (M. Torii).

http://dx.doi.org/10.1016/j.jbi.2015.08.011
1532-0464/© 2015 Elsevier Inc. All rights reserved.

occurrence of a handful of keywords: “nicotine”, “smoker”, “smoke”, “smoking”, “tobacco”, and “cigarette.” In the second step, selected sentences were classified using a k-nearest-neighbors method, and then predicted classes were assigned to the corresponding host documents. The approach by Cohen was similar to Aramaki et al. in that they first identified keywords in a medical record which are occurrences of any of the selected stemmed words: “nicotine”, “smok”, “tob”, “tobac”, “cig”, and “packs.” He called these keywords “hot-spots.” Unlike Aramaki et al., however, he did not select a single sentence per document, but used words near the hot-spots as features for Support Vector Machine (SVM) classifiers [13]. Clark et al. used both a two-step approach, similar to Aramaki et al. and a one-step approach, similar to Cohen.

Hot-spot-based techniques were also successful in the 2011 i2b2/VA/Cincinnati Challenge for sentiment analysis of suicide notes [14]. This task was a multi-label/multi-class classification of sentences from suicide notes, where there were 16 target classes (*Guilt*, *Hopefulness*, *Love*, *Thankfulness*, etc.). In this task, sentences could sometimes be long, but detection of target classes might depend on only a small text segment and often on a very limited vocabulary, *e.g.*, the class *Thankfulness* was mostly associated with the single word “thank(s)” and the class *Love* was associated with the word “love.” Among the best performing systems was Yang et al. [15], who used the hot-spot technique through CRF models. They manually annotated “cue phrases” that are indicative of sentence classes in a development data set, and then trained CRF models to automatically detect the same or similar phrases. These “cue phrases” are essentially the same as hot-spot phrases by Cohen, Aramaki et al. and Clark et al. Given a new sentence, trained CRF models were used to identify cue phrases and, if found, associated classes were assigned to that sentence. Leveraging the CRF models, their system achieved the best results in the 2011 Challenge.

After analyzing the 2014 challenge task, we determined that the task was well suited for a hot-spot-based approach. In designing our system for the 2014 challenge, we leveraged the approaches reported for these past challenge tasks.

2. Materials and methods

2.1. Annotated corpora

Participants in the Track-2 task were provided with two sets of annotated corpora, the Gold corpus and the Complete corpus. Both corpora contain the same source documents that consisted of 790 de-identified clinical notes.

In the Gold corpus, each medical record is provided as an XML file, and target concepts, if reported anywhere in the record, are annotated with XML tags at the record level (*e.g.*, `<CAD time=“during DCT” indicator=“mention” />`). The tags and associated attribute-vales are found in Table 1.

In the Complete corpus, segments of text marked up by three clinicians as evidence annotations are also included. In other words, each concept annotated at the document level in this data set has a reference to the text segment providing the evidence in the record (*e.g.*, `<CAD start=“3575” end=“3579” text=“CAD” time=“during DCT” indicator=“mention”/>`). The corpus is “Complete” in the sense that it includes all the raw annotations by the clinicians, who were requested to record at least the first piece of text that provides supporting evidence in each record in addition to the document level annotation. There could be more than one evidence text segment indicating the same target concept in a record, but they were not exhaustively marked up. Besides, unlike corpora created specifically for training a sequence-labeling model, the annotation boundaries were determined rather arbitrarily.

After the system development period, the challenge participants were provided with an evaluation corpus consisting of 514 medical records, which do not include annotations. Participants applied their system to obtain a result, called a run, and submitted up to three different runs to the organizer for evaluation. Further information regarding these data sets can be found in the overview papers of the 2014 i2b2/UTHealth Challenge [16,17].

2.2. Methods

We built a general text classification system to tackle the diverse sub-tasks in Track 2. Given the success of hot-spot features in prior i2b2 Challenges, we focused on this approach. A general text classifier, a smoking status classifier, and a CRF-based classifier were created that leveraged hot-spot features and also features derived from existing NLP systems. Due to the distinctive properties of the smoking status classification sub-task and the potential benefit of having a standalone tool for that sub-task, an independent module was developed for the sub-task. Different ways of integrating the classification components were explored for the submission runs. The UIMA platform [18] and UIMA compliant tools [19,20] were used to ease the integration.

2.2.1. General text classifier

The general classifier was designed to handle diverse classification sub-tasks in Track 2. In this classification system, each combination of a tag and specific attribute–value pairs was regarded as an independent target category. Then, each of such categories was applicable to some medical records (positive instances) but not to the other records (negative instances). For instance, we considered a tag with specific attribute–value pairs, `<CAD indicator=“mention” time=“before DCT” />`, as an independent target category, and this category was either applicable to a particular record (i.e., the record *does* contain a mention of CAD as an event that the patient previously had) or not (i.e., the record does *not* contain such information). Then, this view defined a binary classification task. That is, in Table 1, each cell with a number entry corresponds to one binary classification task. The number represents exactly the quantity of positive instances of the class, and the negative instances are therefore the remaining complement (i.e., the total number of 790 notes in the training set minus the number of positive instances). For example, in Table 1 (a) Tag: CAD, a cell with the number 260 in row 2 (`time=“before DCT”`) column 1 (`indicator=“mention”`) corresponds to the binary classification task for the aforementioned category, `<CAD indicator=“mention” time=“before DCT” />`, where the number of positive and negative instances are 260 and 530 (=790–260) respectively. The Track-2 task was regarded as a collection of many binary classification tasks. For each of these tasks, we trained a supervised machine learning model that consisted of a classification rule set derived by the RIPPER algorithm [21].

2.2.1.1. General text classifier features.

2.2.1.1.1. Hot-spot features. For each tag (*e.g.*, “CAD”), phrases frequently annotated as evidence text in the Complete corpus (*e.g.*, “coronary artery disease”, “coronary disease”, and “CAD”) were hand-selected and used to identify text segments in a medical record that were immediately relevant to the current classification purpose. Following Cohen [12], we named these selected phrases as hot-spot phrases. Then, selected phrases with similar patterns/concepts were manually grouped together, and a binary feature was defined for the group. For instance, “coronary artery disease”, “coronary disease”, and “coronary heart disease” were grouped together, and if any of these phrases was found in a given medical record, a corresponding feature was set to be 1 or 0 otherwise. The same approach was applied to the MEDICATION tag. For instance,

Please cite this article in press as: M. Torii et al., Risk factor detection for heart disease by applying text analytics in electronic medical records, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.08.011

Download English Version:

https://daneshyari.com/en/article/10355448

Download Persian Version:

https://daneshyari.com/article/10355448

Daneshyari.com