BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 5 2012, pages 743–744
doi:10.1093/bioinformatics/bts030

Data and text mining

Advance Access publication January 19, 2012

medpie: an information extraction package for medical message board posts

A. Benton[1,*], J. H. Holmes[1], S. Hill[2], A. Chung[1] and L. Ungar[3]

[1]University of Pennsylvania School of Medicine, Center for Clinical Epidemiology and Biostatistics, [2]University of Pennsylvania, The Wharton School, Department of Operations and Information Management and [3]University of Pennsylvania School of Engineering and Applied Science, Department of Computer and Information Science, Philadelphia, PA 19104, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: We have developed medpie, a software package for preparing medical message board corpora and extracting patient mentions and statistics for drugs, herbs and adverse effects experienced from them. The package is divided into web-crawling, HTML-cleaning, de-identification and information extraction modules. It also includes a sample controlled vocabulary of drugs, herbs and adverse effect terms.

Availability: http://www.cis.upenn.edu/~ungar/medpie.zip

Dependencies: Python 2.6 or 2.7

Contact: ungar@cis.upenn.edu; adrianb@mail.med.upenn.edu

Received on August 22, 2011; revised on December 21, 2011; accepted on January 11, 2012

1 INTRODUCTION

Medical message boards (MMBs) are sites where patients with similar conditions seek and exchange information by posting messages to threads, collections of messages with similar topics. MMBs can contain a very large number of posts (e.g. over a million messages on http://www.breastcancer.org/), are often freely available and contain patient opinions and experiences that would be potentially useful to clinicians and researchers. However, due to the large number of messages in many MMBs, reviewing each message by hand would be prohibitively slow. In addition, MMB text frequently contains spelling and grammatical errors that may hinder relevant medical information from being automatically extracted as well as personal identifiers such as MMB users' phone numbers, e-mail address and personal names that should be removed before using these data for research. In order to overcome these hurdles, we present medpie, a software package that can be used to generate a de-identified corpus of MMB posts and then extract information from it.

Although researchers have examined message boards and other online communities (Durant *et al*., 2010; Feldman *et al*., 2007), they do not generally make their framework available to others. We have found medpie useful as a first stage in collecting and analyzing MMBs and believe it could facilitate further research of online communities such as MMBs (Benton *et al*., 2011b).

*To whom correspondence should be addressed.

2 DESCRIPTION

The medpie software package consists of four modules written in Python: web-crawling, HTML-cleaning, de-identification and information extraction. The package relies on and includes the CRF++ toolkit (http://crfpp.sourceforge.net/), the BeautifulSoup XML parser (http://www.crummy.com/software/BeautifulSoup/), the porter stemmer implementation from NLTK (http://www.nltk.org/) and a slightly modified Python implementation of Fisher's exact test (http://pypi.python.org/pypi/fisher).

2.1 Web-crawling

The web-crawler can be used to gather message board posts from the internet. This crawler is set to follow and save pages based on whether or not the URL matches a user-defined regular expression. In addition to saving message post pages, the crawler abides by the robots.txt politeness policy and produces a log file listing the pages it has visited as well as saved.

2.2 HTML-cleaning

After the message posts have been saved to the local machine, the HTML has to be 'cleaned' so that only the information relevant to the message posts remains. Sample scripts for cleaning posts from *breastcancer.org* and *healthboards.com* are provided and can be quickly retargeted to other boards. These scripts process the raw HTML files, which are output in a standard XML/JSON format, one thread per file.

2.3 De-identification

The de-identification module removes phone numbers, e-mail addresses, URLs, social security numbers, author usernames and proper names from the subject and body fields of the extracted message posts. The module first identifies e-mail addresses, phone numbers, URLs and social security numbers occurring via regular expression. The text is then tokenized around these regular expression matches and a conditional random field (CRF) trained using the CRF++ toolkit over a 1000 message sample of breast cancer posts is used to identify which tokens are likely to be either proper or usernames and should be removed. Over a sample of 500 messages from the training corpus, the de-identification module correctly removed 98.1% of all proper and usernames in the sample, and over a 500 message sample from an arthritis MMB corpus it

correctly removed 93.8% of all proper and usernames in that sample (Benton *et al*., 2011a). In comparison, MIST, the highest scoring system in the 2006 i2b2 de-identification challenge (Uzuner *et al.* 2007), produced a much lower recall (73.0 and 54.6%, respectively) Although the precision for the system was relatively low (67.4% for the breast cancer corpus), the majority of falsely de-identified tokens was not medically important (e.g. locations, names of famous people).

Note that this module relies on four dictionaries that are included within this package. These dictionaries consist of a list of proper names, approximately 60 000 common English words, the approximately 400 most common English words (stop words) and medical terms. Proper names tend to be de-identified, whereas common, stop and medical words tend to be preserved by this module. These lists can be edited to alter the module's performance. The de-identification module also relies on a 'likelihood parameter' that can be adjusted to vary the likelihood of a token being de-identified as a name. Although recall may be a more relevant metric for evaluating efficacy of de-identification, the 'likelihood parameter' can be adjusted to increase the precision of the system as well.

2.4 Information extraction

This module is used to search for mentions of controlled vocabulary terms within the message board corpus and then return statistics on them. Running this module produces two sets of files, mentions files and statistics files and places them in two separate directories. Mentions files note the file, message number and position that every term in the controlled vocabulary occurs within the corpus, as well as the location of co-occurring pairs of terms. Single term statistics files contain the number of times a given term occurs in the corpus. Term pair statistics files contain term co-occurrence counts within a set number of tokens apart, the likelihood that these terms occurred independently (calculated using the Fisher's exact statistical significance test), and the lift of these two terms (the ratio of observed co-occurrence rate to the co-occurrence rate assuming the terms are independently distributed). This module can be used to get a sense of term prevalence and co-occurrence in the corpus. For example, one can retrieve all statistically significant co-occurrences of drugs and adverse events and then review these messages to determine if mentioned drugs and adverse effects are related.

In addition to these modules, medpie also includes a controlled vocabulary of drug, dietary supplement and event terms.

The drug term vocabulary was scraped from http://www.rxlist.com and contains approximately 5500 terms. The dietary supplement list was compiled by an expert in complementary alternative medicine and contains 507 terms. The event vocabulary consists of approximately 27 000 terms that could either refer to a side effect or indication of a drug. This list was generated by scraping symptom terms from http://www.medicinenet.com/ and adverse effect terms from the Adverse Event Reporting System (AERS) database (http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm). AERS is a database of self-reported drug adverse effects maintained by the US Food and Drug Administration. The vocabulary was augmented with lay synonyms from the Consumer Health Vocabulary (CHV) (Zeng *et al.*, 2006). The CHV is a lexicon with mappings from technical medical terms to lay terms. We found that including these lay terms in our dictionaries improved recall of reported symptoms. Terms that resulted in many false positives were removed from the vocabulary by hand (e.g. boil, shakes). These lists may be altered by the user to change the types of terms that are searched for.

This package includes a small demonstration script that downloads and fully processes 400 messages from *breastcancer.org*. It took ~140 s to execute on a 3.25 GHz dual-core machine, 100 s of which was spent de-identifying the corpus.

3 FUTURE ADDITIONS

A number of additions to the information extraction module are anticipated including modules to identify the speech act of each drug side effect mention discovered and to identify instances of patient drug termination and the motivation behind termination. We also plan to include an open ontology of known drug adverse effects and indications to supplement the package's controlled vocabulary. This ontology could then be used as a filter for drug side effect rules to identify possibly undocumented adverse effects from a particular medication.

Funding: National Library of Medicine (RC1LM010342). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

Conflict of Interest: none declared.

REFERENCES

Benton,A. *et al*. (2011a) A system for de-identifying medical message board text. *BMC Bioinformatics*, **12** (Suppl. 3), S2.

Benton,A. *et al*. (2011b) Identifying potential adverse effects using the web: a new approach to medical hypothesis generation. *J. Biomed. Inform.*, **44**, 989–996.

Durant,K. *et al*. (2010) Social network analysis of an online melanoma discussion group. *AMIA Summits Transl. Sci. Proc.*, **2010**, 6–10.

Feldman,R. *et al*. (2007) Extracting Product Comparisons from Discussion Boards. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*. IEEE Computer Society, Omaha, NE, pp. 469–474.

Uzuner,O. *et al*. (2007) Evaluating the state-of-the-art in automatic de-identification. *JAMIA*, **14**, 550–563.

Zeng,Q. *et al*. (2006) Exploring lexical forms: first-generation consumer health vocabularies. *AMIA Annu. Symp. Proc.*, 2006, 1155.