Automatic abstraction of imaging observations with their characteristics from mammography reports

RECEIVED 26 May 2014
REVISED 11 August 2014
ACCEPTED 4 September 2014
PUBLISHED ONLINE FIRST 28 October 2014

Selen Bozkurt[1], Jafi A Lipson[2], Utku Senol[3], Daniel L Rubin[4,5]

ABSTRACT

Background Radiology reports are usually narrative, unstructured text, a format which hinders the ability to input report contents into decision support systems. In addition, reports often describe multiple lesions, and it is challenging to automatically extract information on each lesion and its relationships to characteristics, anatomic locations, and other information that describes it. The goal of our work is to develop natural language processing (NLP) methods to recognize each lesion in free-text mammography reports and to extract its corresponding relationships, producing a complete information frame for each lesion.

Materials and methods We built an NLP information extraction pipeline in the General Architecture for Text Engineering (GATE) NLP toolkit. Sequential processing modules are executed, producing an output information frame required for a mammography decision support system. Each lesion described in the report is identified by linking it with its anatomic location in the breast. In order to evaluate our system, we selected 300 mammography reports from a hospital report database.

Results The gold standard contained 797 lesions, and our system detected 815 lesions (780 true positives, 35 false positives, and 17 false negatives). The precision of detecting all the imaging observations with their modifiers was 94.9, recall was 90.9, and the F measure was 92.8.

Conclusions Our NLP system extracts each imaging observation and its characteristics from mammography reports. Although our application focuses on the domain of mammography, we believe our approach can generalize to other domains and may narrow the gap between unstructured clinical report text and structured information extraction needed for data mining and decision support.

Key words: Breast Imaging Reporting and Data System (BI-RADS), information extraction, natural language processing, imaging informatics, breast

INTRODUCTION

The interpretation of mammography images is challenged by variability among radiologists in their assessment of the likelihood of malignancy given the abnormalities seen in mammography reports, and methods to improve radiologist performance are needed.[1–3] One approach to reducing radiologist variation is to standardize the vocabulary used in mammography reports. The American College of Radiology (ACR) developed the Breast Imaging-Reporting and Data System (BI-RADS). Terms called 'BI-RADS descriptors' can be used by radiologists to describe breast density, lesion features, impression, and recommendations in mammography reports.[4–6]

While adoption of BI-RADS can help to standardize the vocabulary of reports and reduce the variation in mammography reporting, it is not a decision support system (DSS) for mammography, although others have used it as a key ingredient of DSS.[7,8] For example, several works have demonstrated that statistical machine learning models such as neural networks and Bayesian networks can be built using BI-RADS descriptors and clinical data as inputs.[7–9]

Although preliminary work to develop DSS for mammography using standardized vocabulary to describe the imaging features is promising, few DSS for mammography have been adopted in clinical practice, likely due to the challenge of interfacing DSS with the clinical workflow. DSS for mammography can disrupt the workflow,[7,10–12] since these systems require the radiologist to enter their observations in a separate interface, which duplicates the activity of generating the radiology report. Even if DSS are integrated into a structured reporting system, radiologists generally find it most efficient to produce reports as narrative texts rather than composing reports using structured reporting interfaces. On the other hand, recording

Correspondence to Dr Daniel L Rubin, Department of Radiology, Stanford University, Richard M. Lucas Center, 1201 Welch Road, Office P285, Stanford, CA 94305-5488, USA; E-mail: dlrubin@stanford.edu

© The Author 2014. Published by Oxford University Press on behalf of the American Medical Informatics Association.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

For numbered affiliations see end of article.

Downloaded from https://academic.oup.com/jamia/article/22/2/81/2... by guest on 20 September 2021