Artificial Intelligence in Medicine 66 (2016) 29–39

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Information extraction from multi-institutional radiology reports

Saeed Hassanpour[a,*], Curtis P. Langlotz[b]

[a] Department of Biomedical Data Science, Dartmouth College, 1 Medical Center Drive, Lebanon, NH 03756, United States
[b] Department of Radiology, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, United States

ARTICLE INFO

Article history:
Received 6 April 2015
Received in revised form 22 August 2015
Accepted 24 September 2015

Keywords:
Natural language processing
Information extraction
Discriminative sequence classifier
Radiology report narrative

ABSTRACT

Objectives: The radiology report is the most important source of clinical imaging information. It documents critical information about the patient's health and the radiologist's interpretation of medical findings. It also communicates information to the referring physicians and records that information for future clinical and research use. Although efforts to structure some radiology report information through predefined templates are beginning to bear fruit, a large portion of radiology report information is entered in free text. The free text format is a major obstacle for rapid extraction and subsequent use of information by clinicians, researchers, and healthcare information systems. This difficulty is due to the ambiguity and subtlety of natural language, complexity of described images, and variations among different radiologists and healthcare organizations. As a result, radiology reports are used only once by the clinician who ordered the study and rarely are used again for research and data mining. In this work, machine learning techniques and a large multi-institutional radiology report repository are used to extract the semantics of the radiology report and overcome the barriers to the re-use of radiology report information in clinical research and other healthcare applications.

Material and methods: We describe a machine learning system to annotate radiology reports and extract report contents according to an information model. This information model covers the majority of clinically significant contents in radiology reports and is applicable to a wide variety of radiology study types. Our automated approach uses discriminative sequence classifiers for named-entity recognition to extract and organize clinically significant terms and phrases consistent with the information model. We evaluated our information extraction system on 150 radiology reports from three major healthcare organizations and compared its results to a commonly used non-machine learning information extraction method. We also evaluated the generalizability of our approach across different organizations by training and testing our system on data from different organizations.

Results: Our results show the efficacy of our machine learning approach in extracting the information model's elements (10-fold cross-validation average performance: precision: 87%, recall: 84%, F1 score: 85%) and its superiority and generalizability compared to the common non-machine learning approach (p-value < 0.05).

Conclusions: Our machine learning information extraction approach provides an effective automatic method to annotate and extract clinically significant information from a large collection of free text radiology reports. This information extraction system can help clinicians better understand the radiology reports and prioritize their review process. In addition, the extracted information can be used by researchers to link radiology reports to information from other data sources such as electronic health records and the patient's genome. Extracted information also can facilitate disease surveillance, real-time clinical decision support for the radiologist, and content-based image retrieval.

1. Introduction

Radiology report narrative encompasses critical information about many body parts and health conditions and is a major component of the evidence for clinical diagnosis and disease treatment. In addition, radiology reports provide a rich source of information for disease surveillance, information retrieval, and clinical decision support. However, the free text format of radiology reports and the complexity of natural language make it difficult to extract and re-use report information for clinical care and biomedical research.

Despite this complexity, radiology report narrative mostly follows a common information model consisting of specific semantic elements, such as uncertainty, anatomy, observations, and their modifiers [1]. These common elements capture the essence of

* Corresponding author. Tel.: +1 603 650 1736; fax: +1 877 377 4901.
E-mail addresses: saeed.hassanpour@dartmouth.edu (S. Hassanpour), langlotz@stanford.edu (C.P. Langlotz).

http://dx.doi.org/10.1016/j.artmed.2015.09.007
0933-3657/

report semantics and summarize report information content. Using this information model as a framework for information extraction provides structured details for clinical and research applications and could be generalizable to a wide variety of radiology studies and healthcare organizations. However, identifying and extracting these information model elements is a challenging task due to the ambiguity and subtlety of natural language, the complexity of the described images, and the stylistic variations among radiologists and healthcare organizations.

In this paper, we first present an imaging report information model from an earlier radiology reporting system [1] that defines and summarizes the information content of a radiology report. This information model covers the majority of clinically significant information in radiology reports and is applicable to a wide variety of diagnostic radiology study types. Then we propose an automatic natural language processing (NLP) system to extract clinically significant concepts from the radiology report according to this information model. This system uses a named-entity recognition sequence classifier to identify the information model elements and extract them from the reports. Our approach is applied and evaluated on de-identified radiology reports from three major healthcare organizations: Mayo Clinic (Mayo), MD Anderson Cancer Center (MDA), and Medical College of Wisconsin (MCW).

The main contribution of our work is using existing machine learning techniques to build an information extraction system that can accurately identify significant terms and phrases in radiology reports according to a radiology-specific information model. Our information extraction system yields structured data from the radiology report to link with other clinical and genomic data sources for translational research, information retrieval, disease surveillance, and clinical decision support. The structured data extracted from the radiology report can also improve search of imaging reports for healthcare monitoring and help clinicians and researchers review and understand the reports.

2. Related work

Radiology reports previously have been analyzed using NLP techniques to extract clinically important findings and recommendations [2–4]. The Lexicon Mediated Entropy Reduction (LEXIMER) system extracts and classifies phrases with important findings and recommendations from radiology reports through lexicon-based hierarchical decision trees [3]. In another approach [4], sentences in radiology reports that include clinically important recommendation information were identified though a maximum entropy classifier. Both of these systems provide binary classification (containing or not containing important findings or recommendations) rather than extracting most key clinical concepts using a detailed information model. Our approach enables the re-use of the extracted information for numerous clinical and research purposes, rather than just for the two purposes for which LEXIMER was tailored.

More general NLP techniques previously have been used to classify and extract information from radiology report narrative [5–17]. In earlier work, Medical Language Extraction and Encoding System (MEDLEE) extracted information from Columbia-Presbyterian Medical Center's chest radiology report repository [5]. MEDLEE uses a controlled vocabulary and grammatical rules to translate text to a structured database format. MEDLEE's results were evaluated for 24 clinical conditions based on 150 manually labeled radiology reports [6]. However, in separate studies the authors reported decreases in MEDLEE performance when it was applied to multiple organizations' chest radiology reports [7] and when it was applied to more complex narrative reports from CT and MR head images [8].

In other related work, the Radiology Analysis tool (RADA) was developed to extract key medical concepts and their attributes from radiology reports and to convert them to a structured database format through a specialized glossary of domain concepts, attributes, and predefined grammar rules [9]. Mayo Clinic's Clinical Text Analysis and Knowledge Extraction System (cTAKES) provides a dictionary-based named-entity recognizer to highlight the Unified Medical Language System (UMLS) Metathesaurus terms in text, in addition to other NLP functionalities, such as tokenizing, part of speech tagging, and parsing [10]. As two other widely used UMLS dictionary-based approaches, Health Information Text Extraction (HITEx) from Brigham and Women's Hospital and Harvard Medical School finds UMLS matches to tag principal diagnoses [11] and MetaMap from National Library of Medicine finds UMLS concepts in biomedical literature [12]. A drawback of MEDLEE and other dictionary-based and rule-based annotation and information extraction systems is their limited coverage and generalizability [13]. Building an exhaustive list of terms and rules to model language and extract domain concepts is extremely time consuming. As a result, these dictionary-based and rule-based methods usually suffer from lower recall compared to their precision. In addition, even in the presence of extensive dictionaries and rule bases, the results may be still suboptimal due to the interactions between rules and natural language variations and ambiguity [13].

In related statistical NLP work [14], a statistical dependency parser is combined with controlled vocabulary to capture the relationships between concepts and formalize findings and their properties in a structured format. In another statistical approach, SymText and MPLUS NLP systems combine a controlled terminology, a syntactic context-free grammar parser and Bayesian network-based semantic grammar to code findings in radiology reports [15–17]. Recent related work explored the use of different machine learning methods such as support vector machines and Bayesian networks to classify chest CT scans for invasive fungal and mold diseases at report and patient levels [18,19]. These methods were specialized and evaluated in a limited domains and were not built to extract and summarize the free text information content in multi-institutional radiology reports according to an explicit information model. Our approach improves on the above approaches because a new corpus of annotated reports is not needed to create systems for each new purpose. Also, because our information model is not specific to one type of radiology exam or organization, and it is more generalizable in the domain of radiology.

3. Material and methods

First, we describe the information model that we use to summarize radiology reports, which provides a framework to build and evaluate our information extraction system. Then, we present our radiology report repository and the set of features extracted from the reports in our NLP approach. Our information extraction system is built around a core named-entity recognition method. We propose three different named-entity recognition methods for our information extraction task: (1) dictionary-based method, (2) conditional Markov model (CMM) and (3) conditional random field model (CRF). The first method is commonly used in lexicon-based annotation and information extraction systems and serves as a baseline for two other machine learning methods. Finally, we explain the evaluation mechanism for our system.

3.1. Information model

Our information model provides a framework to summarize radiology reports and to build and evaluate our information extraction system. Our information model's level of detail is optimized to extract and organize the NLP system's results, so they