Journal of Biomedical Informatics 77 (2018) 11–20

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort

Imon Banerjee[a,*], Matthew C. Chen[b], Matthew P. Lungren[b], Daniel L. Rubin[a,b]

[a] Department of Biomedical Data Science, Stanford University, Stanford, CA, United States
[b] Department of Radiology, Stanford University, Stanford, CA, United States

ARTICLE INFO

Keywords:
Information extraction
Word embedding
Pulmonary embolism
Report annotation

ABSTRACT

We proposed an unsupervised hybrid method – Intelligent Word Embedding (IWE) that combines neural embedding method with a semantic dictionary mapping technique for creating a dense vector representation of unstructured radiology reports. We applied IWE to generate embedding of chest CT radiology reports from two healthcare organizations and utilized the vector representations to semi-automate report categorization based on clinically relevant categorization related to the diagnosis of pulmonary embolism (PE). We benchmark the performance against a state-of-the-art rule-based tool, PeFinder and out-of-the-box word2vec. On the Stanford test set, the IWE model achieved average F1 score 0.97, whereas the PeFinder scored 0.9 and the original word2vec scored 0.94. On UPMC dataset, the IWE model's average F1 score was 0.94, when the PeFinder scored 0.92 and word2vec scored 0.85. The IWE model had lowest generalization error with highest F1 scores. Of particular interest, the IWE model (trained on the Stanford dataset) outperformed PeFinder on the UPMC dataset which was used originally to tailor the PeFinder model.

1. Introduction

Radiology is central to modern healthcare, providing detailed clinical information for disease detection, staging and treatment planning while also playing an important role in monitoring and predicting outcomes. Radiology reports are composed of unstructured free-text, and conversion into a computer manageable representation for large scale analysis requires strategies for efficient and automated information extraction. Natural language processing (NLP) tools are designed to convert unstructured text into coded data which may enable automatic identification and extraction of information from radiology text reports for a variety of clinical applications, including diagnostic surveillance, cohort building, quality assessment, labels for computer vision data, and clinical decision support services.

Despite the advantages, NLP remains an underutilized technique for large-volume radiology report data extraction in both research and clinical practice environments due to high development costs and lack of generalizabiltiy of models. Many of the best performing NLP methods are Dictionary-based [1] or Rule-based analysis [2], which, while accurate for a specific task, requires tremendous manual effort to tune the methods for a particular case-study/dataset. Recently, deep learning has provided researchers with tools to create automated classification models without requiring hand-crafted feature engineering which is adapted widely for medical images [3–5]. However, the deep learning methods have yet to show similar performance gains on information extraction from free text radiology reports. A challenge to applying deep learning methods to information extraction in text is modeling ambiguity of free text narrative style for clinical reports, lexical variations, use of ungrammatical and telegraphic phases, and frequent appearance of abbreviations and acronyms.

We propose a hybrid method – Intelligent Word Embedding (IWE) that combines semantic-dictionary mapping and neural embedding technique for creating context-aware dense vector representation of free-text clinical narratives. Our method leverages the benefits of unsupervised learning along with expert-knowledge to tackle the major challenges of information extraction from clinical texts, which include ambiguity of free text narrative style, lexical variations, use of ungrammatical and telegraphic phases, arbitrary ordering of words, and frequent appearance of abbreviations and acronyms. Ideally, the transformation of large volume of free-text radiology reports into dense vectors may serve to unlock rich source of information for solving a variety of critical research challenges, including diagnostic surveillance, cohort building, and clinical decision support services. In this study, we will exploit the embedding created by the IWE method to generate annotation of a large multi-institutional cohorts of chest CT radiology reports based on various level of categorizations of PE.

* Corresponding author.
E-mail addresses: imonb@stanford.edu (I. Banerjee), rubin@stanford.edu (M.C. Chen), mcc17@stanford.edu (M.P. Lungren), mlungren@stanford.edu (D.L. Rubin).

https://doi.org/10.1016/j.jbi.2017.11.012
Received 24 July 2017; Received in revised form 17 November 2017; Accepted 19 November 2017
Available online 23 November 2017

In the targeted case-study, the first important determinant is whether the patient has a PE or not, which informs medical care and treatment decisions; however it is possible that the patient has had prior imaging that diagnosed PE and subsequent imaging may demonstrate an unchanged, diminished, or otherwise chronic PE, in which case medical treatment may change based on whether the PE had responded to prior therapy. Finally it is controversial whether subsegmental PE requires treatment at all, and is not felt to have the same clinical implications as central PE, and thus this category holds valuable importance for clinical decision making [6,7].

We formulated annotation of the radiology reports in terms of three different PE categorical measures (PE present/absent, PE acute/chronic, PE central/subsegmental) as separate classification task. Note that a given report cannot have labels of 'PE acute/chronic' or 'PE subsegmental only/central' without the label of 'PE present'. Our formulation is mainly influenced by the fact that the performance of the 'PE positive' label alone and drawing conclusions in comparison to other NLP classifiers has significant value as the primary clinical state based on the imaging study. The characteristics of 'PE acute' vs 'PE chronic' or 'PE subsegmental' vs 'PE central' location, while important, are each inherently more challenging and have less clinical impact compared to the fundamental disease state and conflating these labeling tasks would provide less information about individual label performance for this exploratory evaluation.

We benchmark the performance of IWE model against a state-of-the-art rule-based solution PeFinder [8] and out-of-the-box word2vec model [9,10] using radiology reports from two major academic institutions: Stanford and University of Pittsburgh medical center. The proposed embedding produced high accuracy (average F1 score Stanford dataset – 0.97, UPMC dataset – 0.94) for three different categorical measures of PE despite the fact that the reports were generated by numerous radiologists of differing clinical training and experience. Besides, the IWE model trained on the Stanford dataset, and used to create embeddings from UPMC dataset, beat the PeFinder model which was originally developed on the UPMC dataset. IWE model also improved upon the out-of-the-box word2vec and showed more generalizability on heterogeneous datasets. We also explored the visualization of vectors in low dimensional space while retaining the local structure of the high-dimensional vectors, to investigate the legitimacy of the semantic and syntactic information of words and documents. In the following sections, we detail the methodology (Section 3), present the results (Section 4) and finally conclude by mentioning core contributions, limitations and future research directions (Section 5).

2. Related works

MedLEE (Medical Language Extraction and Encoding System) in an example of traditional NLP approach in medical domain which relies on controlled vocabulary and grammatical rules in order to convert free-text into a structured database [11,12]. Dang et al. processed 1059 radiology reports with Lexicon Mediated Entropy Reduction (LEXIMER) to identify the reports that include clinically important findings and recommendations for subsequent action [13]. A core limitation of such rule-based systems is that all the kinds of entities and relations need to be pre-specified, and it requires enormous amount of manual effort to initiate such systems if the number of such entities and relations that need to be extracted is significantly large. Moreover, extension of such systems, even for a similar case-study, needs nearly equal amount of manual work.

In addition to traditional dictionary-based and rule-based NLP techniques, various combinations of NLP pipelines and Machine learning methods have been proposed [14,15] that do not demand substantial manual effort and can be retrained without reprogramming for any domain. Sohn et al. used tokenizer combined with machine learning to identify patients with abdominal aortic aneurysms [16]. Nguyen et al. [17] combined traditional supervised learning methods with Active Learning for classification of imaging examinations into reportable and non-reportable cancer cases.

However, the performance of machine learning models heavily depends on finding meaningful vector space projections of the unstructured texts. In most approaches, documents are represented by a simple sparse bag-of-words (BoW) representations which face several challenges in the clinical domain: (i) *scalability* – BoW encode every word in the vocabulary as one-hot-encoded vector, but clinical vocabulary may potentially run into millions; (ii) *semantics of the words* – the vectors corresponding to same contextual words are orthogonal; (iii) *word orderings* – BoW models also don't consider the order of words in the phrase.

There is now an emerging trend with deep learning that adopts a distributed representation of words by constructing a so-called neural embedding of each word or document. The word2vec model introduced by Mikolov et al. [9,10] is the most popular approach for providing semantic word embeddings. One of the biggest challenges with word2vec is how to handle unknown or out-of-vocabulary (OOV) words and morphologically similar words. This can particularly be an issue in domains like medicine where synonyms and related words can be used depending on the preferred style of radiologist, and words may have been used infrequently in a large corpus. If the word2vec model has not encountered a particular word before, it will be forced to use a random vector, which is generally far from its ideal representation. Our proposed method – *Intelligent Word Embedding* (IWE) that can efficiently handle OOV words by combining neural embedding with the semantic dictionary mapping.

3. Material and methods

3.1. Dataset

3.1.1. Cohorts

Stanford dataset – With the approval from the Stanford Institutional Review Borad (IRB), we obtained radiology reports from Stanford medical center for contrast-enhanced CT examinations of the chest performed between January 1, 1998 and January 1, 2016. Using radiology procedure codes, a total of 117,816 CT examinations of the chest with contrast reports were selected for our analysis. All examinations were de-identified in a fully HIPAA-compliant manner and processing of data was approved by the IRB.

Two experienced radiologists performed annotation of total 4512 randomly selected reports. Three binary labels were assigned to individual reports which was defined according to three categorical measures of PE: (1) PE present/absent; (ii) PE acute/chronic; (iii) PE central/subsegmental only. If a PE was definitely present in the report it was annotated as positive for PE present, or else annotated as negative. Chronicity was labeled as either acute or chronic based on the text description. In the setting of acute on chronic, or "mixed" chronicity, the report was labeled as acute to reduce the false negative rate. The "subsegmental only" label was used in cases where the PE was described as subsegmental and did not include more central locations.

Interrater reliability was estimated as Cohen's Kappa Score and the raters were highly consistent for the first two categories of determining PE present and PE Acute with kappa scores of 0.959 and 0.969 respectively. Significant disagreement (kappa score of 0.664) was observed when looking at PE subsegmental label. A senior radiologist resolved all conflicting cases manually for preparing the ground truth labels.

UPMC dataset – We obtained 858 reports from University of Pittsburgh medical center that were originally used to develop PeFinder classifiers. The reports were all de-identified in a fully HIPAA-compliant manner. The annotations were defined according to two categorical measures of PE: (1) PE present/absent; (ii) PE acute/chronic. Three medical students independently annotated the reports with five distinct states and binary annotations for each document were obtained from