Journal of Digital Imaging (2019) 32:685–692
https://doi.org/10.1007/s10278-018-0141-4

Comprehensive Word-Level Classification of Screening Mammography Reports Using a Neural Network Sequence Labeling Approach

Ryan G. Short[1] · John Bralich[2] · Dave Bogaty[2] · Nicholas T. Befera[1]

Published online: 18 October 2018

Abstract

Radiology reports contain a large amount of potentially valuable unstructured data. Recently, neural networks have been employed to perform classification of radiology reports over a few classes at the document level. The success of neural networks in sequence-labeling problems such as named entity recognition and part of speech tagging suggests that they could be used to classify radiology report text with greater granularity. We employed a neural network architecture to comprehensively classify mammography report text at the word level using a sequence labeling approach. Two radiologists devised a comprehensive classification system for screening mammography reports. Each word in each report was manually categorized by a radiologist into one of 33 categories according to the classification system. Tagged words referencing the same finding were grouped into unique sets. We pre-labeled reports with a rule-based algorithm and then manually edited these annotations for 6705 screening mammography reports (25.1%, 66.8%, and 8.1% BI-RADS 0, 1, and 2, respectively). A combined convolutional and recurrent neural network model was used to label words in each sentence of the individual reports. A siamese recurrent neural network was then used to group findings into sets. Performance of the neural network-based method was compared to a rule-based algorithm and a conditional random field (CRF) model. Global accuracy (percentage of documents where all word tags were predicted correctly) and keyword accuracy (percentage of all words that were labeled correctly, excluding words tagged as unimportant) were calculated on an unseen 519 report test set. Two-tailed *t* tests were used to assess differences between algorithm performance, and $p < 0.05$ was used to determine statistical significance. The neural network-based approach showed significantly higher global accuracy compared to both the rule-based algorithm (88.3 vs 57.0%, $p < 0.001$) and the CRF model (88.3% vs. 75.8%, $p < 0.001$). The neural network also showed significantly higher keyword level accuracy compared to the rule-based algorithm (95.5% vs. 80.9% $p < 0.001$) and CRF model (95.5% vs. 76.9%, $p < 0.001$). We demonstrate the potential of neural networks to accurately perform word-level multilabel classification of free text radiology reports across 33 classes, thus showing the utility of a sequence labeling approach to NLP of radiology reports. We found that a neural network classifier outperforms a rule-based algorithm and a CRF classifier for comprehensive multilabel classification of free text screening mammography reports at the word level. By approaching radiology report classification as a sequence-labeling problem, we demonstrate the ability of neural networks to extract data from free text radiology reports at a level of granularity not previously reported.

Keywords Natural language processing · NLP · Deep learning · Radiology reporting

Introduction

The radiology report is an important medical document, serving as the primary mode of communication from radiologist to referring physician, and more recently from radiologist to patient [1, 2]. Radiology reports are stored in electronic health records (EHRs) for documentation of diagnostic imaging interpretation and represent a rich source of unstructured

✉ Ryan G. Short
Ryan.Short@duke.edu

[1] Department of Radiology, Duke University Medical Center, 2301 Erwin Road, Box 3808, Durham, NC 27710, USA

[2] Scanslated, Inc., Durham, NC, USA

medical information. The extraction of information from free-text radiology reports has many potential applications including quality improvement projects, follow-up tracking, clinical research cohort building, and labeling radiology images for computer vision machine learning applications [3]. Unfortunately, extraction of this information often requires substantial manual effort due to the variable structure and content of free-text radiology reports. For the utilization of the medical information contained within radiology reports to be time and cost efficient, automated information extraction techniques are necessary.

Natural language processing (NLP) is one method of obtaining structured data from free text which has been used to automatically derive information from radiology reports [4]. Most applications of NLP to radiology reports have utilized rule-based, simple statistical (e.g., word and phrase frequencies), and/or traditional machine learning methodologies [5–7]. Outside of radiology, new state-of-the-art performance on NLP tasks has been achieved by combining neural networks with distributed representations of words in a vector space [8]. Recent work has shown the efficacy of this methodology in performing document level classification of radiology reports over a few (6–15) classes [9–12].

While document level classification of radiology reports has utility, most reports contain additional valuable information at a more granular level which cannot be extracted by this approach. For example, discrimination between descriptions of similar findings (benign calcifications in the left breast vs. suspicious calcifications in the right breast) cannot be easily accomplished with document level classification. Additionally, radiology reports may describe the same finding in multiple locations, such as a mass in the "findings" section which is then further characterized by an assessment statement in the report impression. Automatic extraction of such detailed information from radiology reports requires a different approach.

The success of sequence labeling for tasks such as named entity recognition (NER) and part of speech (POS) tagging suggests that this approach could be used to classify radiology reports with greater granularity [13, 14]. Recently, Cornegruta et al. employed a recurrent neural network (RNN) in a sequence labeling approach to classify each word in chest radiograph reports into four broad category labels (clinical finding, body location, descriptor, and medical device) [15]. Compared with document level classification, word-level classification of free-text radiology reports may enable extraction of more detailed information with a number of additional useful applications such as the generation of customizable structured radiology reports for referring physicians or novel patient-facing report derivatives [16]. Furthermore, extracted descriptive anatomic information could be used to label radiology images for computer vision machine learning tasks [17].

The purpose of this study was to demonstrate the utility of a sequence labeling approach to word-level classification of screening mammography reports. We compared the performance of three sequence-labeling techniques (a neural network model, a rule-based algorithm, and a conditional random field (CRF) algorithm) in comprehensively classifying free text screening mammography reports at the word level.

Materials and Methods

Corpus and Information Model

We first extracted 6705 screening mammography reports from a database of > 3 million de-identified radiology reports from a single institution. This de-identified database of radiology reports was previously created and was exempted by our institutional review board (IRB).

The automated extraction of information from screening mammography reports was approached as a sequence-labeling problem in which each word in each report was assigned a categorical label. Two radiologists (RGS, NTB) devised a comprehensive word-level classification system for screening mammography reports (Table 1). Each word in each report was manually categorized by a radiologist into one of 33 categories according to the classification system and assigned a corresponding label. Words that were determined to have no meaningful contribution to report content were labeled "none." For example, in the sentence, "there is a 3-cm mass in the right breast which is stable and benign," labeled words (not including words labeled "none") include 3 (size) cm (size), mass (type_1), right (laterality), stable (assesment_3), and benign (assesment_2).

Labeled words referencing the same finding were then sub-labeled as sets and assigned the same set number (e.g., set 1) to indicate that the words were associated and characterized the same finding. In this manner, multiple findings (a mass in the right breast, calcifications in the left breast) in a single report could be accounted for separately. Additionally, set sub-labels also provided a means to discriminate between both multiple instances of a certain type of finding (a mass in the right breast and a mass in the left breast) and multiple descriptions/recommendations of a single finding in different locations within the report (findings: mass in the right breast...impression: right breast mass is suspicious for malignancy). Negated words/phrases (e.g., no architectural distortion) were assigned a label of "none." An example labeled report is shown in Fig. 1.

Rule-Based Algorithm and Manual Annotation

A rule-based NLP framework was devised to automatically annotate screening mammography reports according to the

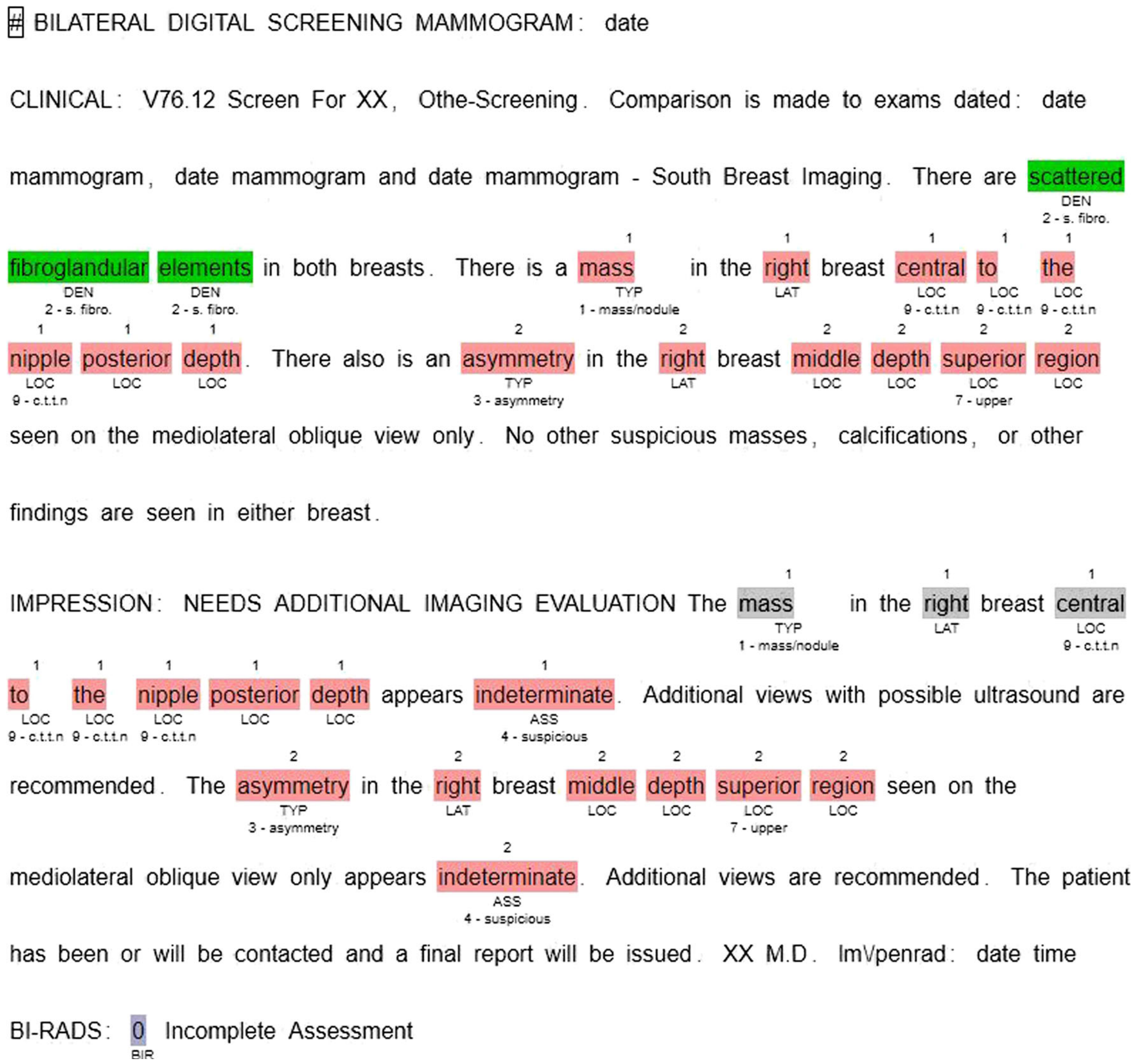
Table 1 Screening mammography report classification scheme

Label	Description
None	Not important
Assessment	Assessment not otherwise specified
Assesment_1	Technical repeat necessary
Assesment_2	Benign
Assesment_3	Stable, unchanged
Assesment_4	Suspicious, requiring diagnostic mammogram
BI-RADS score	BI-RADS score (0, 1, 2)
Density_1	Predominantly fatty
Density_2	Scattered fibroglandular
Density_3	Heterogeneously dense
Density_4	Extremely dense
Descriptor	Finding descriptor (i.e., spiculated, circumscribed, regional, etc.)
Laterality	Laterality (left, right, bilateral)
Location	Location not otherwise specified (anterior, middle, posterior depth)
Location_1	Superior medial location
Location_2	Medial location
Location_3	Inferior medial location
Location_4	Inferior lateral location
Location_5	Lateral location
Location_6	Superior lateral location
Location_7	Superior location
Location_8	Inferior location
Location_9	Central to the nipple, retroareolar location
Size	Size of finding
Type	Type of finding not otherwise specified
Type_1	Mass, nodule
Type_2	Calcification(s)
Type_3	Asymmetry
Type_4	Architectural distortion
Type_5	Lymph node
Type_6	Biopsy clip
Type_7	Post-surgical scar, lumpectomy, mammopexy
Type_8	Implant

information model described above. Briefly, reports were tokenized using the open-source Natural Language Toolkit version 3.2.4 (NLTK) [18]. Using the Python 3.6 package re, we performed concept identification at the word level by using the appropriate regular expression to identify a list of keywords and phrases derived from the Breast Imaging Reporting and Data Systems (BI-RADS) lexicon and supplemented by words and phrases common in our corpus of reports [19–21]. Negated concepts were identified using a customized database of negation phrases which frequently occurred in our corpus of reports. Negated phrases were assigned a “none” label.

The rule-based algorithm processed sentences sequentially in order to assign set labels. A set label was assigned for regex pattern matches classified as location, size, type, assessment, laterality, and descriptor. A set label of one was assigned to the first sentence in the findings section containing words belonging to any of these classes. The next sentence in the findings section containing words from any of these classes was assigned a set label of two. Similarly, the first sentence in the impression section containing words from these classes was assigned a set label of one. The second sentence in the impression section containing words from these classes was assigned a set label of two.

The rule-based algorithm was used to pre-label reports according to the information model described above. Subsequently, one of two radiologists (RGS, NTB) manually edited these annotations for 6705 screening mammography reports (25.1%, 66.8%, and 8.1% BI-RADS 0, 1, and 2, respectively).



BILATERAL DIGITAL SCREENING MAMMOGRAM : date

CLINICAL : V76.12 Screen For XX , Othe-Screening . Comparison is made to exams dated : date

mammogram , date mammogram and date mammogram - South Breast Imaging . There are scattered
DEN
2 - s. fibro.

1 1 1 1 1
fibroglandular elements in both breasts . There is a mass in the right breast central to the
DEN DEN TYP LAT LOC LOC LOC
2 - s. fibro. 2 - s. fibro. 1 - mass/nodule 9 - c.t.t.n 9 - c.t.t.n 9 - c.t.t.n

1 1 1 2 2 2 2 2 2
nipple posterior depth . There also is an asymmetry in the right breast middle depth superior region
LOC LOC LOC TYP LAT LOC LOC LOC LOC
9 - c.t.t.n 3 - asymmetry 7 - upper

seen on the mediolateral oblique view only . No other suspicious masses , calcifications , or other

findings are seen in either breast .

1 1 1
IMPRESSION : NEEDS ADDITIONAL IMAGING EVALUATION The mass in the right breast central
TYP LAT LOC
1 - mass/nodule 9 - c.t.t.n

1 1 1 1 1 1
to the nipple posterior depth appears indeterminate . Additional views with possible ultrasound are
LOC LOC LOC LOC LOC ASS
9 - c.t.t.n 9 - c.t.t.n 9 - c.t.t.n 4 - suspicious

2 2 2 2 2 2
recommended . The asymmetry in the right breast middle depth superior region seen on the
TYP LAT LOC LOC LOC LOC
3 - asymmetry 7 - upper

2
mediolateral oblique view only appears indeterminate . Additional views are recommended . The patient
ASS
4 - suspicious

has been or will be contacted and a final report will be issued . XX M.D . ImVpenrad : date time

BI-RADS : 0 Incomplete Assessment
BIR

Fig. 1 Example labeled report. Words receiving the "none" category label have no highlight or visible annotation. The numbers that are shown above the labeled words represent the sets to which the word belongs

CRF Model

Conditional random fields (CRF) are a class of statistical models that are commonly used for sequence-labeling problems similar to ours including NER and POS tagging [22–24]. We employed a first-order Markov CRF model using the Python-CRF suite implementation [25]. We trained two CRF models, one to predict word classifications, and the other to predict set sub-labels for each labeled word in a given report.

For each word (w_i), features from the three previous words (w_{i-3}, w_{i-2}, w_{i-1}), the current word (w_i), and the next word (w_{i+1}) in the sequence were used for classification. For word classification, the features included a lowercase version of each word, the part of speech tags, a bias, and whether w_{i-1}, w_i, w_{i+1} were digits. In addition to these features, the CRF for set assignment utilized the word classifications for w_{i-1}, w_i, w_{i+1}. In training, we used the truth classifications as features, whereas, in testing, the predicted classifications were used as features. The outputs for this model were a sequence of the 33 possible labels from Table 1, as well as a sequence of the predicted sets in the report.

The models were optimized using the Orthant-Wise Limited-memory Quasi-Newton algorithm with L1 = 0.1 and L2 = 0.01 [26]. The label model was trained for a maximum of 500 iterations and the set model was trained for a maximum of 1000 iterations.

Neural Network-Based Models

Neural networks are also commonly used for sequence-labeling problems and have achieved state-of-the art or near state-of-the art performance in both NER and POS tagging [27]. Based on these successes, we employed a neural network architecture to comprehensively classify free text screening mammography radiology reports at the word level.

First, the training corpus of 5697 screening mammography reports was used to generate word vectors. Word embeddings were generated using the Gensim implementation of word2vec [28]. The continuous bag of words (CBOW) model with a window size of ten was utilized to generate word vectors of dimension $n = 100$.

Neural networks were constructed with Keras 2.0.4 [29].The neural network architecture for the sentence neural network can be seen in Table 2. The input to the network consists of a sequence of embedding integers corresponding to word positions in the word2vec trained embedding matrix. These integers function as a column lookup into the embedding matrix. The embedding matrix weights were initialized based on the word2vec model output and were fine-tuned in training. Full reports were divided into a sequence of sentences by splitting the report at periods. Individual sentences were input into the sentence model. Individual sentences were preprocessed to a uniform length of 55 integers by zero padding the beginning of the sentences. Words that did not occur in the word2vec vocabulary were assigned an embedding integer of zero. The output of the embedding layer was a matrix consisting of the concatenated word vectors in order of the input sequence.

The 2D embedded vector was passed through a convolutional layer with 128 filters of size five (denoted convolutional 1D (128,5)) using stride of one, followed by a bidirectional layer of gated recurrent units (GRU) with 256 units [30]. A time distributed dense layer with sigmoid activation functions was applied to the hidden states corresponding to each input word. The word labels were converted into a list of binary labels for each word. There was an output in the time distributed dense layer for each label. To accommodate for up to three sets of characterized findings in each sentence, three additional set output nodes were combined with the 32 unique class output nodes (the "none" class was not explicitly included in the neural network output) such that each word was described by a vector with length 35.

The sentence network is inherently limited in its ability to correctly associate descriptors of the same finding in two different sentences, which is a common occurrence in radiology reports. In order to overcome this limitation and appropriately group descriptors of the same finding from different sentences, output of the sentence network was passed through a set association neural network (Comparison neural network) as shown in Table 3.

A Siamese architecture, as has previously been used for measuring sentence similarity, was used for the comparison neural network [31]. The comparison neural network had two inputs, A and B, which were the two sets of labeled words to be compared. The two sets of labeled words were zero padded to be uniform length 20 and converted to concatenated word vectors using the same initial embedding layer weights as the sentence neural network. The embedding weights were fine-tuned in training. The 2D vector of embedded words was input to a shared bidirectional GRU layer of 128 units. Element-wise absolute difference was computed between the two final output feature vectors and these differences were passed to a dense layer with sigmoid activation. In training, a value of 0 indicated that different sets were being described, whereas a value of 1 would indicate the same set was being described. In testing, the output score of the dense layer was rounded to 0 or 1 to predict if the same set or different sets were being described.

In testing, an input report was broken up into sentences and the sentence neural network was applied to each sentence individually. The sentences were then processed sequentially. The first sentence with tagged words was assigned a set label of 1 by convention (input A). Using the comparison neural network, the next encountered sentence of tagged words (input B) was compared to set 1. If the comparison neural network predicted input A and input B to be equivalent, input B was assigned the same set as input A (e.g., set 1). If input A and input B were predicted to be different, input B was assigned to a new set (e.g., set 2).

The manually labeled reports were split into training, validation, and test sets with 5,697, 489, and 519 reports, respectively. Using mini-batches of 64 for a total of 100 epochs, each network was optimized using the Adam algorithm to minimize the binary cross-entropy loss function [32]. Validation performance as measured by keyword accuracy was used for model selection.

Evaluation

Global accuracy (percentage of documents where all word tags were predicted correctly) and keyword accuracy (percentage of all words that were labeled correctly, excluding words labeled as "none") were calculated on a previously unseen 519 report test set (32.2%, 60.7%, and 7.1% BI-RADS 0, 1, and 2, respectively). For the neural network approach, accuracy was reported as the mean of three trials. Two-tailed *t* tests were used to assess differences between the three algorithms (rule-based, CRF, neural network) and $p < 0.05$ was used to determine statistical significance.

Results

The global accuracy and keyword accuracy for each of the three methods are presented in Table 4. The neural network was the best performing model, showing higher keyword accuracy (95.5%) compared to the CRF model (77.0%) and the rule-based method (80.9%) ($p < 0.001$). The neural network also showed higher global accuracy (88.3%) compared to the CRF model (75.9%) and the rule-based method (57.0%) ($p < 0.001$). The rule-based method showed higher keyword accuracy compared to the CRF model (80.9% vs. 77.0%,

Table 2 Architecture for the sentence neural network

Input: embedding integers (sentence)
Embedding layer
Convolutional 1D (128,5)
Bidirectional GRU (256)
Time-distributed dense (35)
Output: word-level labels

$p < 0.001$). The CRF model showed higher global accuracy compared to the rule-based method (75.9% vs. 57.0%, $p < 0.001$).

Discussion

Free-text radiology reports are a potential source of valuable, detailed medical information. Recent work shows that a sequence labeling approach may have utility in extracting data from free-text radiology reports at a level of granularity not previously reported [13, 14]. We compared the performance of three sequence-labeling techniques (word-level neural network classifier, rule-based algorithm, CRF classifier) in comprehensive multilabel classification of free text screening mammography reports. Our results demonstrate the superiority of a neural network approach to this sequence-labeling task and highlight the ability of a word-level neural network classifier to overcome limitations encountered with rule-based and CRF techniques.

The rule-based algorithm showed strong performance in keyword accuracy (80.9%) which is due to most keywords in screening mammography reports originating in the limited BI-RADS lexicon. The rule-based algorithm identified and assigned predetermined labels to BI-RADS terms as well as other frequently identified keywords which were used to supplement the BI-RADS lexicon. These words appeared consistently across the corpus of reports.

The rule-based algorithm performed poorly on global classification (57%), likely because many mammography reports contained additional pertinent terms which were not recognized by the algorithm. We anticipated and attempted to address this limitation by supplementing BI-RADS terms with additional words that were common to our training data. We did not, however, incorporate all possible words that could fall under a classification, as this would require a prohibitive number of rules. An additional limitation was that our set assignment assumptions for the rule-based algorithm did not always hold true, as the first set in the in the findings section of a mammography report is not always the first set to appear in the impression section.

Table 3 Architecture for the comparison neural network

Input A: embedding integers A	Input B: embedding integers B
Embedding layer	
Shared bidirectional GRU layer (128)	
GRU features A	GRU features B
Feature-wise absolute difference	
Dense (1) layer	
Output: score range 0–1 (0 different sets, 1 same set)	

Table 4 Performance of the three sequence labeling methods

Method	Global accuracy	Keyword accuracy
Rule-based	57.0%	80.9%
CRF	75.9%	77.0%
Neural network	88.3%	95.5%

The neural network approach improves on the rule-based model due to better flexibility and generalizability. While the rule-based approach is dependent upon a predetermined list of words, the neural network approach utilizes all available training data to learn word labels. Given the number of classes in our data, many more rules would have been necessary for the rule-based model to approach comparable performance to the neural network.

The neural network approach also overcomes limitations encountered with the CRF model. In our comprehensive multilabel classification problem, the ability to model long-range dependencies is essential for correctly assigning set labels. An inherent weakness of CRFs is an inability to model long-range dependencies in a computationally efficient manner [33]. In Fig. 1, for example, successful set assignment for the word “mass” in the impression requires correctly identifying the relationship between this word and the words assigned to set 1 earlier in the findings section of the report. Though the CRF model could accurately predict keyword classification, it was not effective in set assignment. While the overall keyword accuracy of the CRF model was 77.0%, the accuracy was only 66.4% for words that also had a set label compared to 96.1% for words that were not assigned a set label.

The neural network approach improves on the CRF baseline by enabling better modeling of long-range dependencies. By inputting sentences into the sentence neural network, the set association problem is simplified as the network only has to predict labels and sets at the sentence level, rather than at the report level. The comparison neural network is then explicitly optimized for set association across the entire report. The resulting average keyword accuracy for words with set labels is 93.5% with the neural network compared to 66.4% for the CRF model.

Prior studies have shown the efficacy of neural networks in radiology report classification at the document level. For example, Chen et al. and Banerjee et al. classified the impression from chest computed tomography (CT) reports into 6 classes (pulmonary embolism (PE) absent/present, PE acute/chronic, PE central/sub-segmental) [11, 12]. Additionally, Shin et al. and Choski et al. classified head CT reports into 5 classes with 3 sub-classes each [9, 10]. Our work expands on these studies in that we used a sequence learning approach to perform multilabel classification of free-text radiology reports at the word level rather than document level. Document level classification techniques are not a feasible approach for comprehensive screening mammography report classification due to the prohibitively large number of potential report findings.

Applying neural networks to radiology reports to obtain word-level labels has potential valuable downstream applications beyond those possible from prior work. For example, the granular abstracted data from word-level labels could be used as a large source of "truth data" for computer vision tasks [17]. Furthermore, the data could be used to produce customizable structured reports for clinician use or to generate novel patient-facing reports [16].

While this study presents evidence of the efficacy of neural networks for sequence labeling of radiology reports, there are several limitations. First, our study was limited to screening mammography reports, which are typically more structured than other types of radiology reports and make use of a limited lexion (BI-RADS). The performance of neural network sequence labeling on other types of radiology reports is unknown. Future work should explore the application of similar classification systems and neural network architectures to other types of radiology reports. Second, the reports in this work were from a single institutional database and were created using semi-structured reporting software which allows for free dictation. The generalizability of our neural network to mammography reports from other institutions is unknown.

Conclusions

We demonstrate the potential of neural networks to accurately perform word-level multilabel classification of free-text radiology reports across 33 classes, thus showing the viability of a sequence labeling approach to NLP of radiology reports. We found that a neural network classifier outperforms a rule-based algorithm and a CRF classifier for comprehensive multilabel classification of free text screening mammography reports at the word level.

Compliance with Ethical Standards

Conflicts of Interest Ryan G. Short is co-founder and CMO of Scanslated, Inc.

John Bralich is an employee of Scanslated, Inc.

Dave Bogaty is an employee of Scanslated, Inc.

Nicholas T. Befera is co-founder and CEO of Scanslated, Inc.

References

1. Friedman PJ: Radiologic reporting: structure. AJR Am J Roentgenol 140:171–172, 1983
2. Bruno MA, Petscavage-Thomas JM, Mohr MJ, Bell SK, Brown SD: The "Open Letter": Radiologists' Reports in the Era of Patient Web Portals. J Am Coll Radiol 11:863–867, 2014
3. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, Rybicki FJ, Mitsouras D: Natural Language Processing Technologies in Radiology Research and Clinical Applications. Radiographics 36: 176–191, 2016
4. Pons E, Braun LMM, Hunink MGM, Kors JA: Natural Language Processing in Radiology: A Systematic Review. Radiology 279: 329–343, 2016
5. Hassanpour S, Bay G, Langlotz CP: Characterization of Change and Significance for Clinical Findings in Radiology Reports Through Natural Language Processing. J Digit Imaging 30:314–322, 2017
6. Masino AJ, Grundmeier RW, Pennington JW, Germiller JA, Crenshaw, 3rd. EB: Temporal bone radiology report classification using open source machine learning and natural langue processing libraries. BMC Med Inform Decis Mak 16:65, 2016
7. Chen P-H, Zafar H, Galperin-Aizenberg M, Cook T: Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports. J Digit Imaging 31:178–184, 2017. https://doi.org/10.1007/s10278-017-0027-x
8. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J: Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst:3111–3119, 2013
9. Shin B, Chokshi FH, Lee T, Choi JD: Classification of radiology reports using neural attention models. 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp 4363–4370. https://doi.org/10.1109/ijcnn.2017.7966408.
10. Chokshi F, Shin B, Lee T, Lemmon A, Necessary S, Choi J. Natural Langeuage Processing for Classification of Acute, Communicable Findings on Unstructured Head CT Reports: Comparison of Neural Network and Non-Neural Machine Learning Techniques. bioRxiv 2017:173310. https://doi.org/10.1101/173310.
11. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB et al.: Deep Learning to Classify Radiology Free-Text Reports. Radiology 171115, 2017
12. Banerjee I, Chen MC, Lungren MP, Rubin DL: Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort. J Biomed Inform 77:11–20, 2018
13. Wang P, Qian Y, Soong FK, He L, Zhao H. Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network. arXiv [csCL] 2015.
14. Dernoncourt F, Lee JY, Szolovits P. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2017. https://doi.org/10.18653/v1/d17-2017.
15. Cornegruta S, Bakewell R, Withey S, Montana G. Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks. arXiv [csCL] 2016.
16. Short RG, Middleton D, Befera NT, Gondalia R, Tailor TD: Patient-Centered Radiology Reporting: Using Online Crowdsourcing to Assess the Effectiveness of a Web-Based Interactive Radiology Report. J Am Coll Radiol 14:1489–1497, 2017
17. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. arXiv [csCV] 2017.
18. Bird S, Klein E, Loper E. Natural Language Processing with Python. "O'Reilly Media, Inc."; 2009.

19. Welcome to Python.org. Python.org n.d. https://www.python.org/ (accessed March 12, 2018).
20. 6.2. re — Regular expression operations — Python 3.6.4 documentation n.d. https://docs.python.org/3/library/re.html#module-re (accessed March 12, 2018).
21. American College of Radiology. ACR BI-RADS atlas: Breast Imaging Reporting and Data System ; Mammography, Ultrasound, Magnetic Resonance Imaging, Follow-up and Outcome Monitoring, Data Dictionary. 2013.
22. Lafferty JD, Mc Callum A, Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc.; 2001, p. 282–9.
23. Finkel JR, Grenager T, Manning C: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2005, pp. 363–370
24. Silfverberg M, Ruokolainen T, Linden K, Kurimo M: Others. Part-of-speech tagging using conditional random fields: Exploiting sub-label dependencies for improved accuracy, 2014
25. python-crfsuite — python-crfsuite 0.9.5 documentation n.d. https://python-crfsuite.readthedocs.io/en/latest/index.html (accessed March 31, 2018).
26. Andrew G, Gao J. Scalable training ofL1-regularized log-linear models. Proceedings of the 24th international conference on Machine learning - ICML '07, 2007. https://doi.org/10.1145/1273496.1273501.
27. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv [csCL] 2015.
28. Řehůřek R, Sojka P: Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, ELRA, 2010, pp. 45–50
29. Chollet F. Keras 2015.
30. Cho K, van Merrienboer B, Bahdanau D, Bengio Y. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. arXiv [csCL] 2014.
31. Mueller J, Thyagarajan A. Siamese Recurrent Architectures for Learning Sentence Similarity. AAAI 2016.
32. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv [csLG] 2014.
33. Liu F, Baldwin T, Cohn T. Capturing Long-range Contextual Dependencies with Memory-enhanced Conditional Random Fields. arXiv [csCL] 2017.