# Context-Driven Concept Annotation in Radiology Reports: Anatomical Phrase Labeling

**Henghui Zhu[1], Ioannis Ch. Paschalidis, PhD[1], Christopher Hall[2], Amir Tahmasebi, PhD[3]**
**[1]Division of Systems Engineering, Boston University, Brookline, MA, USA**
**[2]Radiology Solutions, Philips Healthcare, Andover, MA, USA**
**[3]Philips Research North America, Cambridge, MA, USA**

**Abstract**

*During a radiology reading session, it is common that the radiologist refers back to the prior history of the patient for comparison. As a result, structuring of radiology report content for seamless, fast, and accurate access is in high demand in Radiology Information Systems (RIS). A common approach for defining a structure is based on the anatomical sites of radiological observations. Nevertheless, the language used for referring to and describing anatomical regions varies quite significantly among radiologists. Conventional approaches relying on ontology-based keyword matching fail to achieve acceptable precision and recall in anatomical phrase labeling in radiology reports due to such variation in language. In this work, a novel context-driven anatomical labeling framework is proposed. The proposed framework consists of two parallel* Recurrent Neural Networks (RNN)*, one for inferring the context of a sentence and the other for word (token)-level labeling. The proposed framework was trained on a large set of radiology reports from a clinical site and evaluated on reports from two other clinical sites. The proposed framework outperformed the state-of-the-art approaches, especially in correctly labeling ambiguous cases.*

**Introduction**

In today's image-driven care practice, radiology reports are commonly used to capture and store clinical observations and the corresponding interpretations by radiologists and to communicate relevant information to the primary care physicians (PCP) and patients. A typical radiology report contains information about abnormalities and disorders observed in the image, which may correspond to multiple organs and anatomical structures. Structuring report content helps with appropriate ingestion of information at every step of the clinical care workflow and eventually, improve the quality of care. As a result, during the last decade, there has been significant interest in the automatic structuring of report content.

Every observation within the medical image(s) has a corresponding anatomical site of reference. Therefore, a promising approach for the automatic structuring of radiology report content is based on identification and sorting of the observations according to their corresponding anatomical site. Nevertheless, the language used for referring to and describing anatomical regions varies significantly among radiologists. Sometimes, the radiologist describes the anatomical site with an elongated phrase (e.g., 'inferior aspect of the glenohumeral joint space, adjacent to the scapular body'). Often, radiologists refer to anatomical regions using custom generated abbreviations (e.g., 'semi. vesc.' referring to 'seminal vesicle'). In other occasions, the transcribed anatomical phrase refers only to a part of an anatomical region without explicitly mentioning the site itself, such as 'wall', 'left lobe', or 'right segment'. This may result in ambiguities to resolve for a machine learning method as it could refer to multiple organs.

Anatomy inference in radiology reports can be performed at word/token, sentence, or document level. The desired level varies based on the application. For example, in order to determine relevant prior imaging studies, a radiologist can take advantage of a solution that could determine the target anatomy of the study at the document (report) level. On the other hand, sentence-level labeling can be used to help with co-referencing problems, such as linking diagnoses to observations that may be found in different sections of the radiology report. Finally, token-level annotations can be used for structuring report context into a searchable database.

The aim of the proposed work is to develop a solution for automatic labeling of anatomical phrases in radiology reports at the token level. Figure 1 demonstrates a snippet from a radiology report with highlighted and labeled anatomical phrases.

Previous efforts for concept labeling can be categorized into three types of approaches: 1) dictionary/ontology lookup;

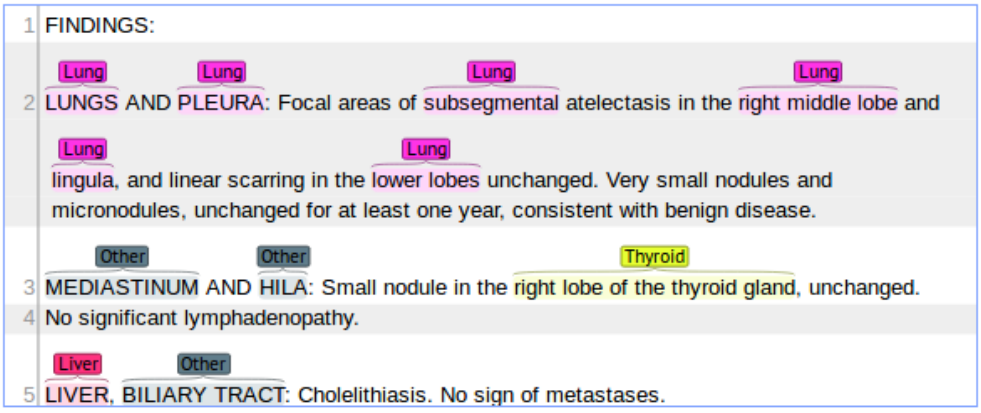



**Figure 1:** A snippet from a radiology report with labeled anatomical phrases. The figure is generated using the BRAT annotation tool.[1]

2) rule/grammars/pattern matching; and 3) data-driven machine learning. As the name implies, dictionary-based approaches rely on existing domain knowledge resources such as ontologies to build keyword dictionaries and the labeling is achieved via string matching. RADA[2] (Radiology Analysis tool), and cTAKES[3] (clinical Text Analysis and Knowledge Extraction System) are among common dictionary-based entity labeling tools. In grammar-based approaches, rules are learned and generated based on repeatable patterns, and morphologies and semantics. In order to provide standard labeling, grammar matching approaches are guided by standard ontologies. MedLEE[4,5] (Medical Language Extraction and Encoding System) and Metamap[6] are among the earliest proposed grammar-based approaches.

Machine learning-based annotators refer to classification models that are learned from clinical text datasets. The anatomical annotation problem is intrinsically related to the Named Entity Recognition (NER). The aim of NER algorithms is to identify and classify target concepts such as anatomical phrases, morphological abnormalities, etc. Deployment of deep learning models for NER-related problems has demonstrated promising results. Deep learning for NER has been successfully proposed for different applications in the medical domain including de-identification,[7] medical events labeling,[8] and clinical concept extraction.[9] Typical architectures proposed for NER applications, especially for sequence labeling problems, include bidirectional *Recurrent Neural Networks (RNN)* models[10,11] and bidirectional RNN-based *Conditional Random Field (CRF)* variants.[12,13]

Despite the promising performance of RNN models for different sequence labeling tasks, capturing the long-term dependency is a remaining shortcoming. This is a major limiting factor for anatomy labeling task. Consider the following sentence: 'The right lobe of the lung is clear, but the 5mm ground glass nodule in the upper left lobe may require further follow up'. It is straightforward to determine the anatomical label for 'right lobe of lung' as it contains the organ name; however, in order to determine the label for 'upper left lobe' at the end of the sentence, the anatomy cue existing in 'right lobe of lung' in the beginning of the sentence should be taken into account. Such distance relation would be difficult for an RNN model to learn through its memory-based architecture.

In this work, we propose a context-driven approach for automatic labeling and normalization of anatomical phrases in radiology reports. The proposed framework consists of two parallel RNNs. Given a sentence from a report, the first RNN model is used to determine the anatomical context at the sentence-level. The second RNN generates token-specific feature vectors. Finally, the sentence-level feature vector and the token-specific feature vectors are combined to derive the most appropriate anatomical label for each token in the sentence. The proposed framework enforces decision-making using the context of the sentence for anatomy annotation.

The main contributions of the proposed work is as follows:

- A context-driven deep learning approach is proposed for anatomical phrases labeling in radiology reports. Given

the complexity of the NER tasks in radiology informatics due to the specific language found in radiology reports, context is proven to play a significant role in achieving an acceptable performance. To the best of our knowledge, this is the first work proposing to utilize context to improve a NER tool performance for a clinical application.

- The proposed study highlights an important finding regarding memory capacity of the RNN models: In this study, we demonstrated that despite the expectations, the long short-term memory architecture of RNN models is not sufficient to guarantee context learning even within the same sentence and as a result, we propose an additional RNN encoder dedicated to solely capture and learn the context.

### Clinical Data

Radiology reports from two different clinical sites, University of Washington (UW) and University of Chicago (UC), as well as reports from a publicly available database, referred to as MIMIC-III[14], were used for training and testing. Radiology reports from UW and UC were collected with Institutional Review Board (IRB) approvals. All reports were de-identified by offsetting dates with randomly generated numbers. All other HIPAA protected patient health information including name, date of birth and address were removed. The following describes the distribution of the data:

- **Word embedding model training:** 1,567,581 radiology reports for examinations performed between January 1, 2010, and February 28, 2017, from the radiology information system at UW were extracted, processed and used for training the word embedding model.
- **Anatomy labeling training set:** 560 radiology reports from UW (referred to as UW560), excluded from the word embedding training set, were randomly extracted and manually labeled using eight anatomical labels as follows: *Brain*, *Breast*, *Kidney*, *Liver*, *Lung*, *Prostate*, *Thyroid*, and *Other* referring to all other anatomical sites (e.g., *Spine*, *Heart*). The remaining non-anatomical tokens in the report were automatically labeled as *Null*. The reason for selecting such seven class labels is to simplify the manual ground truth generation for a relatively large corpus to be used for training and validation as the manual labeling for the whole human anatomy would require a significant amount of time and effort. 70% of this set was randomly selected for training (UW560-70%) and the remaining 30% was used as the development set to determine the optimal set of parameters (UW560-30%).
- **Anatomy labeling test set:** In order to avoid any bias due to the training corpus, 200 radiology reports from two different clinical sites were considered and manually labeled using the same labeling schema as mentioned above and used as the test set: 100 reports randomly selected from the UC database (UC100); and 100 reports randomly selected from the MIMIC-III database (MIMIC-III100).

### Ground Truth

Manual labeling was performed by human annotators (not domain experts) based on the eight aforementioned labels using BRAT annotation tool[1]. The annotators referred to SNMOED CT[15] ontology for determining the classes of anatomy named entities. The UW560 training set was labeled by two annotators. No Inter-Annotator Agreement (IAA) measure was calculated for this round. The testing set was manually labeled by four human annotators. Identified phrases with more than two disagreement between annotators in terms of selecting a class label were further reviewed by a radiologist for determining the appropriate label. The overall IAA in terms of average Kappa[16] over each annotator pair was 90.1%, 87.4% for UC100, and MIMIC-III100, respectively.

### Methods

Figure 2 demonstrates an overview of the proposed framework. The following sections detail each step as depicted in the figure.

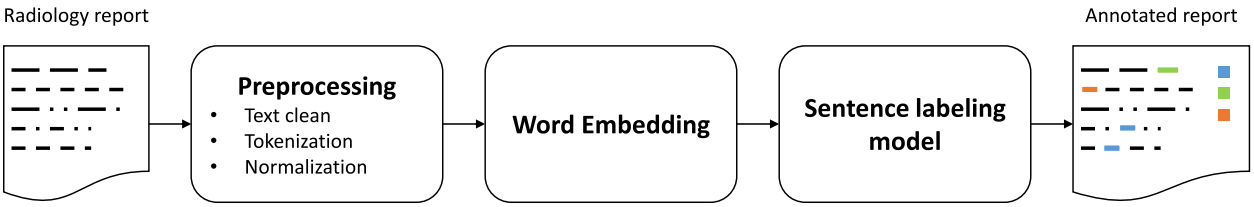



**Figure 2:** A overview of the proposed framework for anatomical phrase labeling in radiology reports.

### Preprocessing

A sentence parsing is firstly performed using spaCy[1] to extract sentence boundaries in a given radiology report. Each extracted sentence is then processed for removing special characters (e.g., HTML tags), unnecessary white space, and new lines. Next, tokenization is applied to partition each text string into individual words. Finally, a normalization step is considered for specific types such as case (lowercase alphabetic), dates and times (make all identical), and numerical values (replace each labeled token with '9'). The whole framework is implemented in Python.

### Word Embedding

The input to RNN architecture is the word embedding representation of tokens. Word embedding refers to the transformation of a string representation of tokens into low-dimensional real-number vectors derived through a set of unsupervised language modeling and feature learning techniques. A few of the most popular word embedding generation approaches are continuous bag of words (CBOW), skip-grams (SG)[17], GloVe[18], and Swivel (SW)[19]. Word embeddings are desired for their ability to capture similarity between words with respect to semantic relationships and are purely learned from unlabeled data. Word embeddings have been used as input feature vectors for many deep learning-based sequence labeling approaches.

Through an exhaustive search, four different word embedding approaches (CBOW, SG, GloVe, and SW) and the corresponding hyper-parameters (including embedding vector size, and context window size) were considered and compared for word embedding model creation. A large corpus of radiology reports from UW (as described before) was used as the training set. The best performing model was determined as: Method: SG, window size: 10, vector size: 500[20]. Since the word embedding model is trained on the radiology reports, we observe very few Out-Of-Vocabulary (OOV) cases in the training and test corpora.

### Sequence Labeling Model

Figure 3 depicts the architecture of the proposed context-driven sequence labeling. The aim of the sequence labeling is to provide a label for every token within a sentence.

To overcome one of the known limitations of the RNN architecture, referred to as fading memory, this paper proposed a framework consisting of two parallel RNN architectures: The first RNN (shown on the left side of Figure 3) is considered to derive a feature vector capturing the anatomy of reference based on the context of the target sentence. A second RNN (shown on the right side of Figure 3) is considered to derive token-specific features. The input to both RNNs is the word embedding vectors for every token within a sentence. The word embedding vectors are provided from the skip-gram model as described before.

The output from the last hidden state in the left RNN (Figure 3), referred to as the *sentence encoder*, provides a feature vector that captures the context information (in terms of the anatomy of reference) at the sentence-level. Since a sentence may contain multiple anatomical phrases referring to different anatomical sites (e.g., 'No abnormality is observed within the urinary system including prostate and kidney.'), a multi-label classification schema is considered for the sentence encoder. The sentence encoder feature vector is passed to a fully-connected layer with nine hidden neurons corresponding to nine class labels and a sigmoid activation function. The output of the fully-connected layer

[1] https://spacy.io/

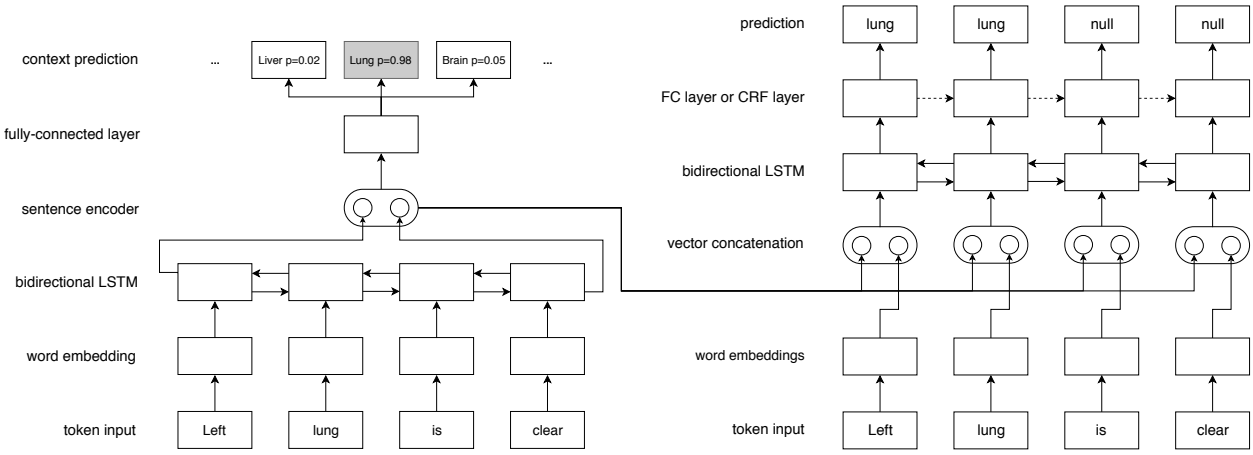



**Figure 3:** The proposed context-driven sequence labeling architecture. The sentence encoder, shown in the left, provides the prediction of the context at the sentence-level. The feature vector from the sentence encoder (shown with green-color circle) is used as an additional feature for the sequence labeling model as shown in right. For the given example, the sentence-level label is determined as *Lung* as shown with gray-color box.

provides the probability of the sentence referring to each of the nine class labels. A cross-entropy loss function is considered for the training by enforcing the sentence encoder to contain anatomical context information.

The second RNN, shown on the right side of Figure 3, has a similar architecture as the first one except that the output of the RNN at every time step (corresponding to every token in the sentence) is used to derive the feature vector for every token in the sentence. Token feature vectors from the second RNN are then concatenated with the sentence encoder feature vector to derive a new higher dimension vector per token. The output of the RNN is then passed to a fully-connected layer with a *softmax* activation or a linear-chain CRF-layer[12] for predicting the label. For a softmax layer, a cross-entropy loss function was considered for the classification of each token.

### Results

The number of token occurrence for each class label within the training and test sets are given in Table 1. As expected, *Null* class has the highest frequency of occurrence followed by the *Other* class. Also, it can be observed that among the seven anatomical classes, *Lung* has the highest and *Prostate* has the lowest frequency. It is expected that such imbalanced representation of classes pose an impact on the word embedding vector representation as well as sequence labeling using RNN architecture as both approaches are shown to be heavily biased based on the training data.

**Table 1:** Number of token occurrence per anatomical class.

| Anatomy Class | UW560 | UC100 | MIMIC-III100 |
|---|---|---|---|
| *Brain* | 1,317 | 235 | 356 |
| *Breast* | 557 | 198 | 2 |
| *Kidney* | 1,678 | 120 | 131 |
| *Liver* | 2,427 | 129 | 185 |
| *Lung* | 3,836 | 291 | 425 |
| *Prostate* | 220 | 6 | 7 |
| *Thyroid* | 269 | 32 | 0 |
| *Other* | 23,636 | 2,385 | 1,784 |
| *Null* | 152,195 | 21,622 | 19,108 |

A number of the most well-known conventional and the state-of-the-art sequence labeling techniques were considered and implemented for the evaluation and comparison as described below. cTAKES[3], as the most common NLP tool

for clinical applications, is used as the baseline system for the anatomy labeling task. In current implementation of cTAKES, the labeling is achieved by first running the clinical pipeline to label all appropriate entities at phrase level, followed by matching *AnatomicalSiteMention* entities with a pre-defined set of SNOMED CT codes (see Table 2) corresponding to the seven anatomical classes in the following fashion: A relation tree is firstly built based on the SNOMED CT relationship *is_a* with roots defined as in Table 2. The depth of the tree is limited to be 7. The phrases matching any of the SNOMED CT codes within the trees are labeled with the corresponding root node anatomy, while the non-matching phrases are labeled as *Other*. The final output is labels at token-level.

**Table 2:** SNOMED CT root node per anatomical class.

| Anatomy | SNOMED CT root node |
|---|---|
| *Brain* | 1101003 |
| *Breast* | 76752008 |
| *Kidney* | 304582006 |
| *Liver* | 10200004 |
| *Lung* | 699594007, 400141005 |
| *Prostate* | 41216001 |
| *Thyroid* | 297261005 |

In addition, vanilla bi-directional RNN models as well as bi-directional RNN-CRF models[12] are implemented and compared with our model. The RNN architectures are similar to the model in Figure 3, except for the sentence encoder part. Two options were considered for the input to the model: 1) passing entire report content at once; 2) pass one sentence at a time. The reason for such choices is to investigate whether providing more content (full report content) helps in achieving more accurate token labeling.

The Adam optimizer[21] was used with step size of $10^{-3}$ for 300 epochs. The dropout approach[22] was considered to prevent overfitting. The following parameters and the corresponding range of values were considered and compared to determine the best performing model. This process was performed for all baseline models as well as the proposed model.

- RNN cell type: LSTM[23], and GRU[11];
- RNN depth: 1, and 2;
- Number of hidden states: 32, 64, 128, and 256;
- Keep probability in the dropout layer: 0.1, 0.3, 0.5, 0.7, and 0.9.

The F1-score was used as the metric for evaluating and comparing the performance of different classification models. The F1-score is calculated as the micro-average of the F1-scores from all eight anatomical classes (excluding the *Null* class) at the token-level. In total, 80 different models (two possibilities of RNN model, two choices for RNN depth, four choices for the number of hidden states and five choices for keep probability in the dropout layers) were trained using the UW560-70% dataset. UW560-30% was used to determine the best performing architecture and the corresponding set of hyperparameters. The best performing model was determined as: one-layer LSTM model with 256 hidden neurons and keep probability of 0.3 for the dropout layer.

Table 3 summarizes the performance of different approaches in terms of F1-score tested on two test corpora, UC100 and MIMIC-III100. As can be observed from the table, the proposed context-driven approach outperformed vanilla RNN models by an average of 1.2% in F1-score. As expected, adding the context information improved the performance of the sequence labeling. Precision, Recall, and F1-score of the best performing model (context-based bidirectional RNN) for each anatomical label are given in Table 4. As can be observed from the table, the proposed model consistently yields high performance for all seven anatomical classes and across two different corpora except for *Prostate*. The lower performance on the *Prostate* class could be due to a low occurrence in the training corpus. Finally, comparing performance between two testing corpora, it can be observed that UC100 yielded higher F1-score

**Table 3:** Comparison of Precision (P), Recall (R), and F1-score (%) among different models. Metrics are defined at token-level.

| Model | UC100 P | UC100 R | UC100 F1 | MIMIC-III100 P | MIMIC-III100 R | MIMIC-III100 F1 | Combined F1 |
|---|---|---|---|---|---|---|---|
| cTAKES[3] | 86.2% | 63.2% | 73.0% | 82.8% | 48.3% | 61.0% | 67.7% |
| Bi-directional RNN (sentence-level) | 84.9% | 88.9% | 86.9% | 86.6% | 83.5% | 85.0% | 86.0% |
| Bi-directional RNN (report-level) | 86.0% | 89.8% | 87.8% | 89.2% | 82.7% | 85.8% | 86.9% |
| Bi-directional RNN-CRF[12] (sentence-level) | 85.4% | 89.0% | 87.2% | 85.2% | 83.6v | 84.5% | 85.9% |
| Bi-directional RNN-CRF[12] (report-level) | 86.7% | 90.5% | 88.6% | 88.9% | 84.0% | 86.3% | 87.6% |
| Our proposed bi-directional RNN | 88.3% | **92.7%** | **90.5%** | **88.9%** | **84.8%** | **86.9%** | **88.8%** |
| Our proposed bi-directional RNN-CRF | **88.6%** | 92.3% | 90.4% | 88.4% | 83.3% | 85.8% | 88.3% |

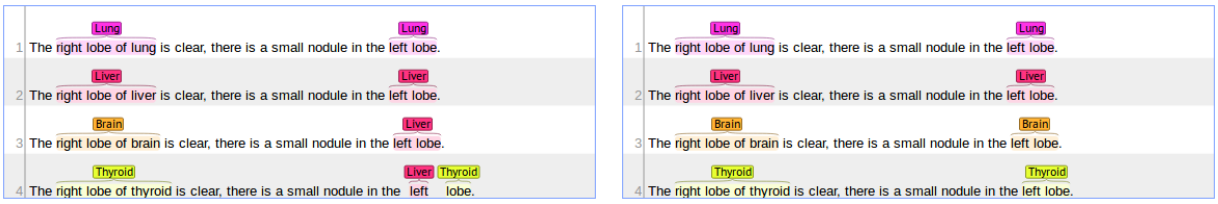
**Table 4:** Per-class Precision (P), Recall (R), and F1-score (F1) of the best model (our proposed context-based bi-directional RNN) on two test datasets: UC100 and MIMIC-III100.

| Class label | UC100 P | UC100 R | UC100 F1 | MIMIC-III100 P | MIMIC-III100 R | MIMIC-III100 F1 |
|---|---|---|---|---|---|---|
| *Brain* | 83.3% | 80.4% | 81.8% | 93.5% | 68.5% | 79.1% |
| *Breast* | 97.2% | 86.4% | 91.4% | 100.0% | 100.0% | 100.0% |
| *Kidney* | 97.0% | 80.8% | 88.2% | 95.5% | 80.2% | 87.1% |
| *Liver* | 86.5% | 94.6% | 90.4% | 88.5% | 87.0% | 87.7% |
| *Lung* | 86.6% | 97.9% | 91.9% | 92.5% | 89.4% | 90.9% |
| *Prostate* | 42.9% | 100.0% | 60.0% | 75.0% | 85.7% | 80.0% |
| *Thyroid* | 90.3% | 87.5% | 88.9% | N/A | N/A | N/A |
| *Other* | 88.1% | 92.5% | 90.2% | 86.8% | 85.0% | 85.9% |
| micro-average | 88.3% | 92.7% | 90.5% | 88.9% | 84.9% | 86.8% |

compared to MIMIC-III100. This could be due to the fact that MIMIC-III reports are only from the intensive care unit; however, reports of the UC100 are from a random mixture of different departments, which is more similar to the training corpus.

### Discussion and Conclusions

Recently, RNN-based architectures have demonstrated promising performance for sequence labeling tasks mainly due to the capability to use its internal memory structure to take into account past and future data in the decision-making. For the token labeling task, this means taking into account the information from words occurring before and after of the target word. Nevertheless, if the desired context cannot be inferred from the immediately surrounding context, an RNN model may not yield the correct labeling.




**(a)** Best vanilla RNN-CRF model. **(b)** Our best context-based bi-directional RNN model.

**Figure 4:** Comparing the output of the best vanilla RNN model and our proposed context-driven RNN model shown for a specific anatomical phrase, 'left lobe', but with different anatomy of reference. Snapshots are generated using the BRAT annotation tool.

In this work, we propose an RNN-based approach for token annotation in radiology reports utilizing sentence-level context to influence the labeling based on a wider context than the immediately surrounding words. This is achieved by adding a second RNN architecture to exclusively learn the context of a given sentence and incorporate that information into the labeling task. To better demonstrate the capability of the proposed framework in utilizing context for labeling, consider examples shown in Figure 4. As can be observed from the figure, given a fixed phrase 'left lobe' but different reference anatomies, the proposed approach is able to yield correct labeling by taking the surrounding context into account. On the other hand, the vanilla RNN-based approach yields incorrect labels. The incorrect labeling by vanilla RNN approach (i.e., *Liver*) could be due to the fact that 'left lobe' occurs most frequently with *Liver* anatomy within the training corpus. As a result, in lack of explicit presentation of an anatomical reference within the vicinity of the target word, the algorithm tends to choose the most frequent label based on the co-occurrence within the training data. Here is another example: "IMPRESSION: Mildly heterogeneous and increased hepatic echotexture, suggestive of parenchymal dysfunction." The only model that yielded the correct label, *Liver*, for *parenchymal* was our proposed context-based RNN. All other models label this term as *Kidney*, which again could be due to the co-occurrence frequency of *parenchymal* and *Kidney* within the training corpus.

One of the major shortcomings of the proposed framework is the lack of a strategy to deal with imbalanced training data. As can be seen from Table 1, there is a large difference between *Other* and *Null* and the rest of the anatomical classes. This may have a direct impact on the performance of the proposed classifier. One approach to deal with such imbalanced data is to use a regularization term in the optimizer to take into account the frequency of occurrence as a weight. Another limitation of the proposed framework becomes obvious when the target sentence context does not contain relevant and sufficient information to help with narrowing down the decision-making. For example, consider the following sentence: 'There is a small lesion in the left lobe.' The target anatomical phrase is 'left lobe'. Even a human expert cannot determine the correct anatomical label without taking more context into account. If the preceding sentences are also provided as: 'Brain: The right lobe is clear.', it becomes clear that the target sentence is also referring to the *Brain* anatomy. As hinted by the example, one approach to overcome such limitation is to use a few sentences before and after the target sentence as the input to the context-based RNN architecture to help with creating the most appropriate context related features.

Overall, the proposed framework demonstrated promising performance for anatomical phrase labeling in radiology reports for a specific list of anatomical class labels. In the future, we are planning to extend the scope of the classification to the whole human anatomy rather than seven classes by utilizing one-shot or few-shot learning algorithms[24,25].

## References


1. Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics, 2012.

2. David B Johnson, Ricky K Taira, Alfonso F Cardenas, and Denise R Aberle. Extracting information from free text radiology reports. *International Journal on Digital Libraries*, 1(3):297–308, 1997.

3. Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

4. George Hripcsak, Carol Friedman, Philip O Alderson, William DuMouchel, Stephen B Johnson, and Paul D Clayton. Unlocking clinical data from narrative reports: a study of natural language processing. *Annals of internal medicine*, 122(9):681–688, 1995.

5. Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 1994.

6. Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

7. Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606, 2017.

8. Abhyuday N Jagannatha and Hong Yu. Structured prediction models for rnn based sequence labeling in clinical text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 856. NIH Public Access, 2016.

9. Willie Boag, Elena Sergeeva, Saurabh Kulshreshtha, Peter Szolovits, Anna Rumshisky, and Tristan Naumann. Cliner 2.0: Accessible and accurate clinical concept extraction. *arXiv preprint arXiv:1803.02245*, 2018.

10. Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

11. Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

12. Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

13. Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

14. Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

15. Jeremy Rogers and Olivier Bodenreider. SNOMED CT: Browsing the Browsers. In *KR-MED*, pages 30–36, 2008.

16. Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

17. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

18. Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

19. Noam Shazeer, Ryan Doherty, Colin Evans, and Chris Waterson. Swivel: Improving embeddings by noticing what's missing. *arXiv preprint arXiv:1602.02215*, 2016.

20. Amir M Tahmasebi, Henghui Zhu, Gabriel Mankovich, Peter Prinsen, Prescott Klassen, Sam Pilato, Rob van Ommering, Pritesh Patel, Martin L Gunn, and Paul Chang. Automatic normalization of anatomical phrases in radiology reports using unsupervised learning. *Journal of digital imaging*, pages 1–13, 2018.

21. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

22. Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027, 2016.

23. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

24. Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In *AAAI*, volume 16, pages 2786–2792, 2016.

25. Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, 2016.