PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Context-based bidirectional-LSTM model for sequence labeling in clinical reports

Henghui Zhu, Ioannis Ch. Paschalidis, Amir M. Tahmasebi

Henghui Zhu, Ioannis Ch. Paschalidis, Amir M. Tahmasebi, "Context-based bidirectional-LSTM model for sequence labeling in clinical reports," Proc. SPIE 10954, Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications, 109540J (15 March 2019); doi: 10.1117/12.2512103

SPIE.

Event: SPIE Medical Imaging, 2019, San Diego, California, United States

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 22 Apr 2019 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Context-based Bidirectional-LSTM Model for Sequence Labeling in Clinical Reports

Henghui Zhu[a], Ioannis Ch. Paschalidis[a], and Amir M. Tahmasebi[b]

[a]Division of Systems Engineering, Boston University, Brookline, MA, USA
[b]Philips Research North America, Cambridge, MA, USA

ABSTRACT

Recurrent Neural Network (RNN) models have been widely used for sequence labeling applications in different domains. This paper presents an RNN-based sequence labeling approach with the ability to learn long-term labeling dependencies. The proposed model has been successfully used for a Named Entity Recognition challenge in healthcare: anatomical phrase labeling in radiology reports. The model was trained on labeled data from a radiology report corpus and tested on two independent datasets. The proposed model achieved promising performance in comparison with other state-of-the-art context-driven sequence labeling approaches.

Keywords: Named Entity Recognition, Sequence Labeling, Nature Language Processing

1. DESCRIPTION OF PURPOSE

A typical radiology report contains descriptions of radiological observations and abnormalities corresponding to different anatomical structures. Structuring of radiology reports content helps improve the quality of care by streamlining communication of the clinical information between the patient's care team through the clinical workflow. With increase in usage of medical imaging services and subsequently an increase in number of radiology reports, numerous efforts have been dedicated towards the automation of structuring the radiology reports content. Most radiologists transcribe reports by ordering content based on the anatomy of reference. Therefore, a common approach for automatic structuring of the report content is to build a Named Entity Recognition (NER) model for detecting anatomical phrases within every sentence, and furthermore, normalizing detected phrases based on predefined categories (i.e., *Lung*, *Liver*, *Brain*, etc.).

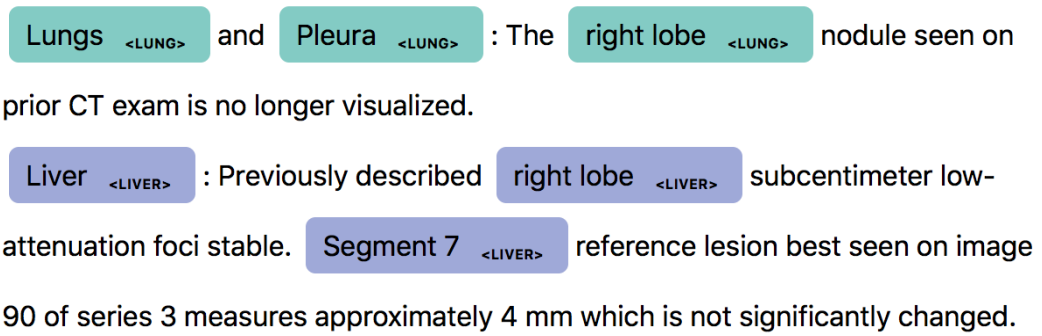


Figure 1: A snippet from a radiology report with highlighted anatomical phrases and corresponding anatomical labels.

Supervised and semi-supervised learning approaches have been proposed for NER problems in different domains utilizing probabilistic graphical models such as Conditional Random Fields (CRFs), and Hidden Markov

Further author information: (Send correspondence to H. Z.)
H. Z.: E-mail: henghuiz@bu.edu
I. P.: E-mail: yannisp@bu.edu
T. A. : E-mail: Amir.Tahmasebi@philips.com

Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications,
edited by Po-Hao Chen, Peter R. Bak, Proc. of SPIE Vol. 10954, 109540J
© 2019 SPIE · CCC code: 1605-7422/19/$18 · doi: 10.1117/12.2512103

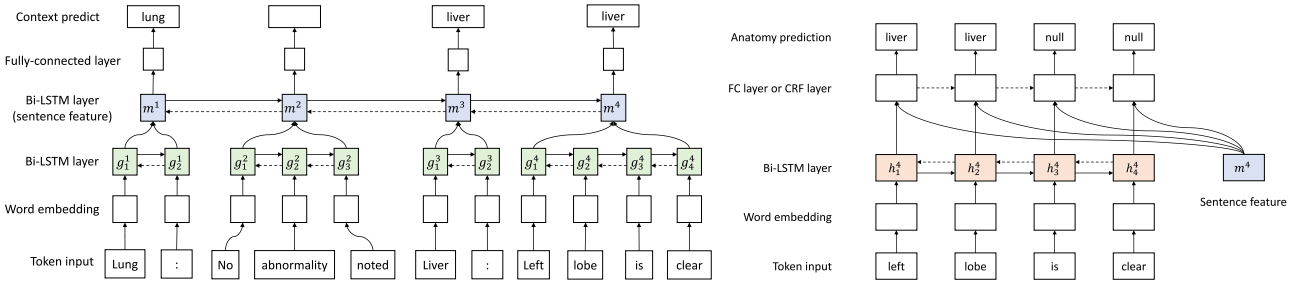
Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 22 Apr 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Models (HMMs). More recently, Neural Network models such as Recurrent Neural Networks (RNN) has been successfully proposed for NER applications.[4,6,12] A major challenge in developing a NER model for the proposed application compared to use cases from other domains is the short- and long-term dependency of the labeling task on the context of the sequence. Consider the example in Figure 1. In this example, several anatomical phrases are highlighted and labeled with the appropriate anatomical labels. Three different labeling scenarios can be seen in this example. First scenario corresponds to phrases that contain a noun with exact match with one of the predefined labels (e.g., "*Liver*", "*Lung*)". Second and third scenarios correspond to phrases without the exact anatomical labels such as "*Segment 7*" and "*Right Lobe*". In case of "*Segment 7*", as the second scenario, the labeling task is straightforward as "*Segment*" is an anatomical phrase that is uniquely used to refer to a division within "*Liver*". In third scenario, no such one to one mapping exist. For example, "*Right/Left Lobe*" is an anatomical phrase that is used to refer to a division of different anatomical structures such as "*Liver*", "*Lung*", "*Brain*", and "*Thyroid*". In example shown in Figure 1, "*Right lobe*" appears in two different sentences referring to two different anatomies: (*Lung*, and *Liver*). Assigning the correct anatomical label to such anatomical phrases with ambiguities as given in this example, is not feasible without taking the surrounding context into account. Most classical approaches including dictionary-based,[9] rule-based,[1] and bag of word-based machine learning,[10] are unable to yield correct labeling for cases with long term labeling dependency such as the one shown in Figure 1. Jagannatha and Hong used an approximate skip-chain CRF model for learning the long-term dependency of labels.[7] Language models can be used to provide features that can capture the contextual meaning of the words in a semi-supervised way. Some related works include context2vec[13] and Elmo,[15] which provide additional information to word embedding for NLP tasks.

In this paper, we present a context-based RNN architecture for sequence labeling in clinical text that is capable of learning short- and long-term labeling dependencies. The proposed approach utilizes supervised learning of the context of each sentence in a document through a bi-directional LSTM (bi-LSTM) architecture while taking into account labels from surrounding sentences. The hypothesis is that the combination of features derived from the sentence context with lexical features derived for each token results in a boost in performance for the anatomical phrase labeling task. Utilizing such short- and long-term learning dependencies, we demonstrate that our proposed approach outperforms state-of-the-art methods for labeling cases with one-to-many labeling possibilities.

2. METHODS

The aim of this work is to develop an automatic approach for extracting and labeling anatomical phrases in radiology reports. Our proposed model consists of two parallel LSTM architectures to derive and combine a token-level feature with a sentence-level feature for classifying the anatomical label of a word in a sequence. The sentence-level feature provides contextual information of a sentence and the token-level features capture lexical features. A diagram of the proposed architecture is shown in Figure 2.



(a) Sentence Presentation (b) Token Presentation

Figure 2: An overview of the proposed architecture.

2.1 Token-Level Features

This work uses Long Short-Term Memory (LSTM),[5] a typical RNN architecture, for sequential modeling. Admittedly, there are some other types of RNN models that are frequently used in language modeling such as Gate

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 22 Apr 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Recurrent Unit (GRU).[3] Nevertheless, the difference in performance is not substantial and therefore, in this work, we will focus on LSTM architecture. In detail, an LSTM model is defined as follows:

$$\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{W}_i\left[\mathbf{x}'_t, \mathbf{h}'_{t-1}\right]' + \mathbf{b}_i\right), \\
\mathbf{j}_t &= \tanh\left(\mathbf{W}_j\left[\mathbf{x}'_t, \mathbf{h}'_{t-1}\right]' + \mathbf{b}_j\right), \\
\mathbf{f}_t &= \sigma\left(\mathbf{W}_f\left[\mathbf{x}'_t, \mathbf{h}'_{t-1}\right]' + \mathbf{b}_f\right), \\
\mathbf{c}_t &= \mathbf{c}_{t-1} \odot \mathbf{f}_t + \mathbf{i}_t \odot \mathbf{j}_t, \\
\mathbf{o}_t &= \sigma\left(\mathbf{W}_o\left[\mathbf{x}'_t, \mathbf{h}'_{t-1}\right]' + \mathbf{b}_o\right), \\
\mathbf{h}_t &= o_t \odot \tanh(c_t),
\end{aligned}$$

where $\mathbf{x}$ is the input of the LSTM, $\mathbf{h}$ and $\mathbf{c}$ denote the state of the LSTM, $\mathbf{W}_i$, $\mathbf{W}_j$, $\mathbf{W}_f$, $\mathbf{W}_o$ are trainable weight matrices and $\mathbf{b}_i$, $\mathbf{b}_j$, $\mathbf{b}_f$, $\mathbf{b}_o$ are trainable biases. $\sigma(\cdot)$ represents an element-wise sigmoid function, $\tanh(\cdot)$ is the hyperbolic tangent function, and $\odot$ denotes element-wise multiplication of two vectors.

Assume $\mathbf{x}_t^i$ represents the word embedding vector of a token t within sentence i of a given report. We use a bidirectional LSTM (bi-LSTM) model to derive a vector representation of tokens within a sentence as follows (Figure 2b):

$$\overrightarrow{\mathbf{h}}_t^i = \text{LSTM}(\overrightarrow{\mathbf{h}}_{t-1}, \mathbf{x}_t^i), \tag{1}$$

$$\overleftarrow{\mathbf{h}}_t^i = \text{LSTM}(\overleftarrow{\mathbf{h}}_{t+1}, \mathbf{x}_t^i), \tag{2}$$

$\mathbf{h}_t^i = [\overrightarrow{\mathbf{h}}_t^i, \overleftarrow{\mathbf{h}}_t^i]$ is the token-specific feature vector.

2.2 Sentence-Level Features

We use an additional bi-LSTM to derive the feature representation at the sentence level as follows:

$$\overrightarrow{\mathbf{g}}_t^i = \text{LSTM}(\overrightarrow{\mathbf{g}}_{t-1}, \mathbf{x}_t^i), \tag{3}$$

$$\overleftarrow{\mathbf{g}}_t^i = \text{LSTM}(\overleftarrow{\mathbf{g}}_{t+1}, \mathbf{x}_t^i), \tag{4}$$

$\mathbf{s}_i = [\overrightarrow{\mathbf{g}}_{L_i}^i, \overleftarrow{\mathbf{g}}_0^i]$ represents the feature vector of the sentence i, which is the concatenation of the output of the forward LSTM at the last token and the output of the backward LSTM at the first token, and L_i is the length of the sentence i. The constructed feature vector tends to learn short-term labeling dependencies.

In radiology reports, it is commonly noted that the radiologist transcribes multiple observations with respect to a specific anatomy without explicitly referring to the anatomy of reference in every sentence. Consider the example in Figure 1. To determine the anatomical label of the sentence "Previously described right lobe subcentimeter low-attenuation foci stable.", the labeling information from the previous sentences needs to be taken into account. In order to learn long-term labeling dependencies that extends beyond the context of the sentence containing the target token, we consider applying an additional bi-LSTM to derive the contextual representation of a sentence influenced by labeling context from the surrounding sentences. In detail,

$$\overrightarrow{\mathbf{m}}^i = \text{LSTM}(\overrightarrow{\mathbf{m}}^{i-1}, \mathbf{s}_i), \tag{5}$$

$$\overleftarrow{\mathbf{m}}^i = \text{LSTM}(\overleftarrow{\mathbf{m}}^{i+1}, \mathbf{s}_i), \tag{6}$$

$\mathbf{m}^i = [\overrightarrow{\mathbf{m}}^i, \overleftarrow{\mathbf{m}}^i]$ is used as a sentence-level feature.

As the last step, we add a fully-connected layer to the sentence level presentation for predicting the anatomical context of the sentence. In detail, for kth anatomy, z_i^k is defined as an indicator function such that $z_i^k = 1$ if there is any token related to kth anatomy appearing in sentence i, and $z_i^k = 0$ otherwise. We use the loss function:

$$L_{\text{context}} = \sum_{i=1}^{L}\sum_{k=1}^{8} z_i \log \sigma(\mathbf{w}_k^T \mathbf{m}^i), \tag{7}$$

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 22 Apr 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

where $\mathbf{w}_k$ is the vector in the fully-connected layer corresponding to the anatomy k and σ is a sigmoid function.

2.3 Context-based LSTM Model

The feature vector for token t is generated by concatenating the token-level feature h_t^i and the sentence-level feature $\mathbf{m}^i$: $\mathbf{p}_t^i = [\mathbf{h}_t^i, \mathbf{m}^i]$. This feature is used for classifying the anatomical label of the token. Two types of loss functions are considered:

1. Softmax loss: An additional fully-connected layer with a softmax activation is used to predict the anatomical label of the token. The loss in sequence labeling is defined as the cross-entropy between the label and its prediction.
2. CRF loss: Similar to,[6] a CRF layer is used for inferring the most likely anatomical label of the token. The loss in this case is the CRF loss in a sequence.

3. DATASETS

Tow large radiology report corpora were used in this study. 1,567,581 reports from University of Washington (UW); and 66,099 radiology reports from University of Chicago (UC). Reports were collected with Institutional Review Board (IRB) approvals. All reports were de-identified by offsetting dates by randomly generated numbers. All other HIPAA patient health information including name, date of birth and address were removed. In order to test the robustness of the proposed approach, a third independent radiology corpus was considered from a publicly available database, referred to as MIMICIII.[8] The data distribution is as follows:

- **UW:** All reports from the UW corpus were used for training a word embedding model.
- **UC500:** 500 reports were randomly selected from the UC corpus. This set was used for the training and development.
- **UC100:** Another set of 100 reports were randomly selected from the UC corpus. This set was used as a part of the testing set.
- **MIMICIII100:** 100 radiology reports were randomly selected from the MIMICIII dataset. These were also included as a part of the testing set.

Table 1: Numbers of occurrence per anatomical class within the training and testing corpora.

	UC500	UC100	MIMICIII100
brain	661	235	356
breast	515	198	2
kidney	810	120	131
liver	669	129	185
lung	1519	291	425
prostate	45	6	7
thyroid	131	32	0
other	10717	2385	1784

All selected reports were manually labeled for anatomical phrases by human annotators (UC500: one annotator; UC100, and MIMICIII100: four annotators). The Inter Annotator Agreement (IAA) measure in terms of Kappa[2] for UC100 and MIMICIII100 were 90.1% and 87.4%, respectively. Eight anatomical labels were considered in this study: *Brain*, *Breast*, *Kidney*, *Liver*, *Lung*, *Prostate*, *Thyroid*, and *Other* referring to all other anatomies. The number of occurrence for each anatomical label within the three corpora are given in Table 1.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 22 Apr 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

4. RESULTS

Sentence parsing was performed using Spacy.* Next, special characters (e.g., HTML tags), unnecessary white space, and new lines were removed. This step was followed by tokenization into individual words. Finally, normalization was performed for lowercasing, unifying the dates, times, and numericals.

Table 2: Precision (P), Recall (R), and F1-score on two test corpora.

Methods	UC100			MIMICIII100		
	P	R	F1	P	R	F1
bi-LSTM-sentence	89.9%	92.8%	91.3%	91.9%	83.8%	87.7%
bi-LSTM-report	89.0%	92.3%	90.6%	90.7%	84.2%	87.3%
bi-LSTM-CRF-sentence[6]	89.1%	92.3%	90.7%	90.8%	84.2%	87.4%
bi-LSTM-CRF-report[6]	90.5%	93.0%	91.7%	91.9%	83.7%	87.6%
ASC CRF-sentence[7]	90.3%	92.6%	91.4%	91.3%	84.6%	87.8%
ASC CRF-report[7]	90.4%	92.3%	91.3%	91.8%	83.9%	87.7%
Our approach with softmax-loss	90.6%	94.0%	**92.3%**	91.4%	85.0%	88.1%
Our approach with CRF-loss	89.4%	94.5%	91.9%	90.1%	86.5%	**88.3%**

The UW corpus was used for training the word embedding model using Skip-gram approach[14] with windows size 10 and vector size 500. We observed only a few Out-Of-Vocabulary (OOV) tokens in the training and testing datasets. The OOV tokens were mapped to the 'unk' token.

The output layer consists of nine classes: eight anatomical labels as defined before and the 'null' label for non-anatomical tokens. Performance was measured and compared across different models based on Precision, Recall, and F1-score and was calculated using data from all class labels except for 'null'.

A few state-of-the-art methods were considered for the comparison as follows: 1) standard bi-LSTM; 2) bi-LSTM with CRF proposed by;[6] and 3) approximate skip-chain (ASC) CRF proposed by.[7] Two types of inputs were considered: 1) sentence-level: the input is only a sentence at a time; and 2) report-level: the whole report is used as the input.

The training set, UC500, was randomly split to 80% and 20% for training and development, respectively. The following hyperparameters and the corresponding range of values were tested to determine the best performing LSTM model for the given task using the development set for all the models: LSTM hidden units: 32, 64, 128, and 256; LSTM depth: 1 and 2; dropout ratio: 0.3, 0.5, and 0.7. We used the Adam optimizer[11] with learning rate of 1e-3 and mini-batches of size 16 for sentence-level input and eight for report-level input for the training. All models were trained for 300 epochs with early stopping based on the development set results.

Precision, Recall, and F1-score for different models are shown in Table 2. Our proposed context-based LSTM outperformed all other models on both testing corpora based on F1-scores. Specifically, our proposed model yielded significantly higher recall compared to other models. This suggests that it is beneficial to include the sentence-level features for building an anatomy sequence label model.

5. NEW OR BREAKTHROUGH WORK TO BE PRESENTED

To the best of our knowledge, for the first time, we propose an RNN-based approach for token annotation in radiology reports utilizing the surrounding sentence-level context to influence the decision making for labeling.

6. CONCLUSION

In this work, we presented a context-based LSTM model for addressing a sequence labeling problem in healthcare: anatomical phrase labeling in radiology reports. Our proposed model is specifically well suited for NER problems that require short- and long-term dependency consideration in labeling decisions. This is achieved by deriving and

*https://spacy.io/

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 22 Apr 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

combining token- and sentence-level features. Our validation results on reports from multiple sites demonstrated that our proposed model outperforms other similar context-driven state-of-the-art sequence labeling approaches suggesting that it is beneficial to include the sentence-level features for building an anatomy sequence label model.

REFERENCES

[1] 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.

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

[3] 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.

[4] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. Deep semantic role labeling: What works and whats next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 473–483, 2017.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] Zhenfei Ju, Jian Wang, and Fei Zhu. Named entity recognition from biomedical text using svm. In *Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on*, pages 1–4. IEEE, 2011.

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

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

[13] Oren Melamud, Jacob Goldberger, and Ido Dagan. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, 2016.

[14] 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.

[15] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 22 Apr 2019
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use