Original Paper

# An Attention Model With Transfer Embeddings to Classify Pneumonia-Related Bilingual Imaging Reports: Algorithm Development and Validation

Hyung Park[1*], MD; Min Song[2*], PhD; Eun Byul Lee[2], BA; Bo Kyung Seo[2], BA; Chang Min Choi[1,3], MD

[1]Department of Pulmonary and Critical Care Medicine, Asan Medical Center, Seoul, Republic of Korea

[2]Yonsei University, Seoul, Republic of Korea

[3]Department of Oncology, Asan Medical Center, Seoul, Republic of Korea

[*]these authors contributed equally

**Corresponding Author:**
Chang Min Choi, MD
Department of Pulmonary and Critical Care Medicine
Asan Medical Center
Olympic-ro 43-gil
Seoul, 05505
Republic of Korea
Phone: 82 2 3010 5902
Fax: 82 2 3010 6968
Email: ccm9607@gmail.com

## Abstract

**Background:** In the analysis of electronic health records, proper labeling of outcomes is mandatory. To obtain proper information from radiologic reports, several studies were conducted to classify radiologic reports using deep learning. However, the classification of pneumonia in bilingual radiologic reports has not been conducted previously.

**Objective:** The aim of this research was to classify radiologic reports into pneumonia or no pneumonia using a deep learning method.

**Methods:** A data set of radiology reports for chest computed tomography and chest x-rays of surgical patients from January 2008 to January 2018 in the Asan Medical Center in Korea was retrospectively analyzed. The classification performance of our long short-term memory (LSTM)–Attention model was compared with various deep learning and machine learning methods. The area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve, sensitivity, specificity, accuracy, and F1 score for the models were compared.

**Results:** A total of 5450 radiologic reports were included that contained at least one pneumonia-related word. In the test set (n=1090), our proposed model showed 91.01% (992/1090) accuracy (AUROCs for negative, positive, and obscure were 0.98, 0.97, and 0.90, respectively). The top 3 performances of the models were based on FastText or LSTM. The convolutional neural network–based model showed a lower accuracy 73.03% (796/1090) than the other 2 algorithms. The classification of negative results had an F1 score of 0.96, whereas the classification of positive and uncertain results showed a lower performance (positive F1 score 0.83; uncertain F1 score 0.62). In the extra-validation set, our model showed 80.0% (642/803) accuracy (AUROCs for negative, positive, and obscure were 0.92, 0.96, and 0.84, respectively).

**Conclusions:** Our method showed excellent performance in classifying pneumonia in bilingual radiologic reports. The method could enrich the research on pneumonia by obtaining exact outcomes from electronic health data.

***(JMIR Med Inform 2021;9(5):e24803)*** doi: 10.2196/24803

**KEYWORDS**

deep learning; natural language process; attention; clinical data; pneumonia; classification; medical imaging; electronic health record; machine learning; model

## Introduction

Electronic health records (EHRs) have become increasingly incorporated into clinical practices in hospitals over the past few decades [1]. EHR data are voluminous and can be used as real-world evidence if they are analyzed with proper methods [2]. However, the data are not collected for research purposes [2], and several rule-based methods are used to extract particular outcomes from the data set. There have been numerous studies where analyses were performed using EHR data with labels such as *sepsis* defined by rule-based outcomes [3-6]. However, defining outcomes other than laboratory findings is difficult because the data are unstructured and written as natural language. For this reason, a previous study that used the outcome *pneumonia* defined pneumonia by its International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) code [7,8]. However, the use of ICD codes as a label does not contain temporal information, such as the exact time of diagnosis during hospital admission, and it is hard to perform time series analysis with this limited information.

Although medical imaging reports contain a great deal of information regarding diagnosis and clinical features, it is hard to analyze the information because they are formatted as unstructured free text and are variably written depending on the radiologist.[9] For this reason, medical imaging reports are rarely used as outcomes in big data analysis [10]. However, as long as pneumonia can be identified in radiologic reports, other important information, such as the time of onset and the presence of pneumonia during admission, can also be derived. Moreover, labeled data are essential in deep learning because the analysis requires millions of observations to reach acceptable performance levels [11].

As of 2018, 43 studies using natural language processing for the identification of chronic diseases in EHRs had been published, and only recently have there been more studies conducted on this topic using deep learning [12]. Especially in deep learning, convolutional neural network (CNN)–based models have shown significant accuracy in extracting pulmonary embolism [10] and pulmonary infection from medical reports [1]. The model can be used to classify diagnosis from whole medical records even when they are written in the Chinese language [13], and a recurrent neural network–based model has been used for classifying stroke and identifying its location [14]. However, the use of bilingual clinical reports is common for EHRs in non–English-speaking countries.

The purpose of our study was to classify reports of pneumonia consisting of findings derived during the pre- and postoperative period of a major surgery that were written as bilingual texts (English and Korean). We compared the performance of traditional models with deep learning models, with the latter showing excellent performance in previous studies, and identified the best performing model as an attention-based bidirectional long short-term memory (Bi-LSTM) model neural network.

## Methods

### Clinical Data

We retrospectively included radiology reports for chest computed tomography (CT) and chest x-rays of surgical patients from January 2008 to January 2018 in the Asan Medical Center in Korea. The patients had undergone upper abdominal and thoracic surgeries, as coded by the ICD-9-CM. Detailed criteria for the surgery are described in Multimedia Appendix 1.

The radiology reports consist of chest CT and chest x-rays (posteroanterior and anteroposterior) that are extracted by radiology procedure codes. The chest x-ray reports have no structured format and only contain descriptions. The chest CT reports consist of the short history of the patients, the findings, and a conclusion; however, the format varies depending on the writing style of the radiologist. The conclusions in around half of the chest CT reports were omitted due to the different writing style of the radiologists. Therefore, we used only the findings of chest CT and the descriptions of chest x-rays to classify the labels, and all the annotation was based solely on the description of each report.

Usually, the pneumonia incidence in surgical patients is around 1%, suggesting that reports of pneumonia are rare. To overcome the imbalance of the positive and negative data sets, we only included radiologic reports that contained pneumonia-related words. The words representing pneumonia were as follows: “pneumoni-,” “consolid-,” “infiltra-,” “bronchiole-,” “hazi-,” “hazzi-,” “opacit-,” and “GGO”.

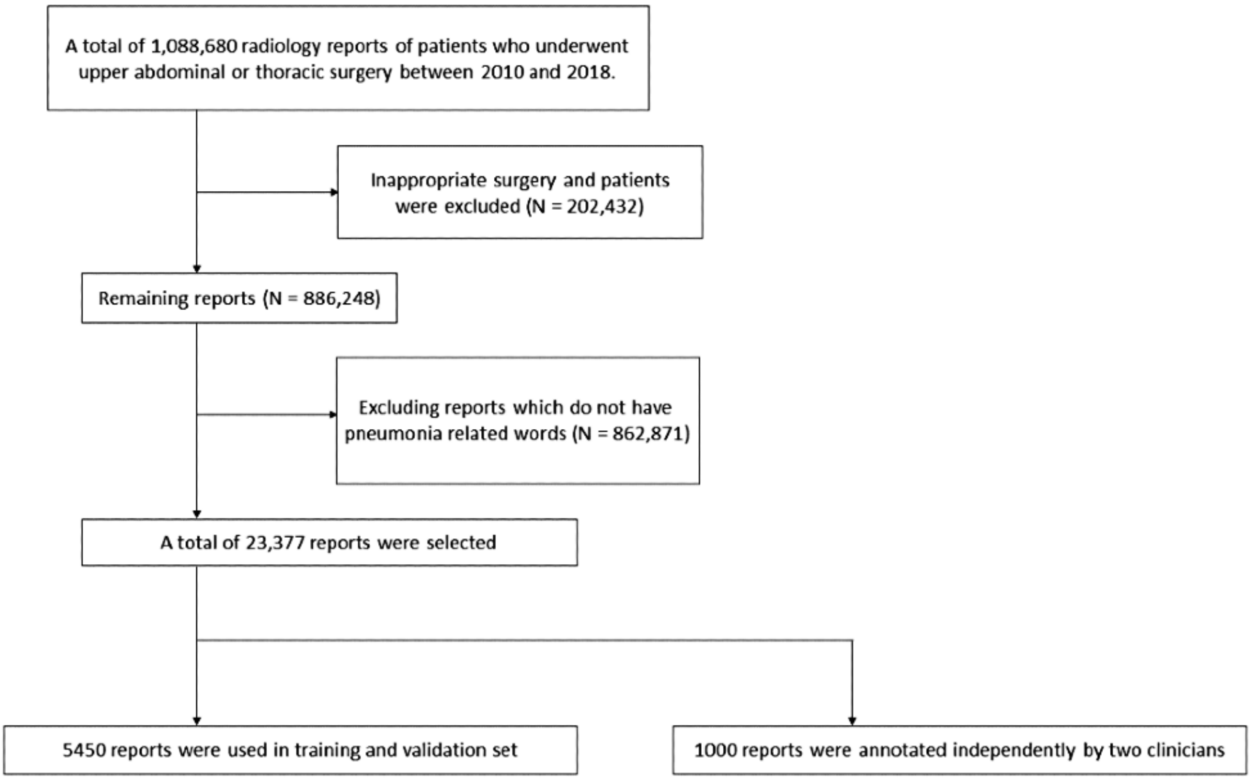
From a total of 1,088,680 radiology reports, 886,248 were included after reports with inappropriate surgical procedures were excluded. The detailed inclusion criteria of the appropriate procedures have been described in a previous study [3]. After extracting the pneumonia-related words, 23,377 reports were included.

### Report Annotation

Among the 23,377 reports, a total of 5450 annotated reports were used to train our model. A clinician annotated the 5450 reports and used them for training and validation. After training the model, 2 different clinicians, who worked independently from the first clinician, annotated another 1000 reports for an extra-validation set (Figure 1).

All document-level annotations by clinicians included 3 categories for pneumonia: negative, positive, and unclear (obscure). The positive pneumonia reports included postoperative infection reports and did not contain reports for noninfectious diseases, such as organizing pneumonia or interstitial lung disease, because the label was required to represent pneumonia as a perioperative complication. The excluded reports were labeled as negative reports. It was observed that 895 reports were pneumonia positive, 4005 reports were pneumonia negative, and 550 reports were obscure results. In the extra-validation set, 2 clinicians independently labeled the radiologic reports on the basis of the clinical importance of the findings. To overcome the human error of the 2 clinicians, the consensus label of the 2 clinicians was regarded as the reference standard. An interrater reliability (k score) was calculated by Cohen κ value.

**Figure 1.** Radiologic reports flowchart.


A total of 1,088,680 radiology reports of patients who underwent upper abdominal or thoracic surgery between 2010 and 2018.
Inappropriate surgery and patients were excluded (N = 202,432)
Remaining reports (N = 886,248)
Excluding reports which do not have pneumonia related words (N = 862,871)
A total of 23,377 reports were selected
5450 reports were used in training and validation set
1000 reports were annotated independently by two clinicians

## Ethics Approval

This study was approved by the ethics committee of the Asan Medical Center (approval no. 2018-1122), and the need to obtain informed consent was waived because of the retrospective observational nature of the study. The clinical data that were extracted using the Asan Biomedical Research Environment system were indexed by deidentified encrypted patient ID numbers so that the researchers would not be able identify the patients [15,16].

## Proposed Approach

As most of the verbs and adjectives in clinical reports are written in Korean, and most of nouns (usually the names of the diseases) are written in English, we had to consider 2 different languages. Therefore, we proposed a new method for a bilingual clinical data set based on the classification algorithm of combining substring and translation embeddings (Kor2Eng) with an attention-based Bi-LSTM neural network (LSTM-Attention). Multimedia Appendix 1 Figure S4 shows the architecture of our proposed model.

The proposed method includes 3 steps: (1) text preprocessing; (2) word representation, which is composed of substring and Korean-to-English (Kor2Eng) embeddings; and (3) training of the classification model.

Our data set, which is a description of x-ray and CT, is composed of a mix of Korean and English sentences. Therefore, specific preprocessing is required before the statements are fed into the classification model. The detailed methods for text preprocessing and training are described in Multimedia Appendix 1.

## Kor2Eng Transfer Embedding

Training word vectors require a considerable amount of data and time. Therefore, we applied embeddings by training them independently on monolingual data and pretraining them with Wikipedia data. However, due to the characteristics of data, the text of the clinical notes was a mixture of English and Korean. If a monolingual embedding were to be used for this data, one side of the information would be lost. To reduce the loss of information, we used a translation method that converts the vector of Korean words into the vector of English words with similar meanings. The unsupervised method of translating the source language into the target language was proposed by Lample et al [17]. In this method, the process of learning a mapping occurs between the 2 sets of embedding in the shared space. We trained the subword embedding model to learn Korean-to-English mapping using the unsupervised method without any parallel data.

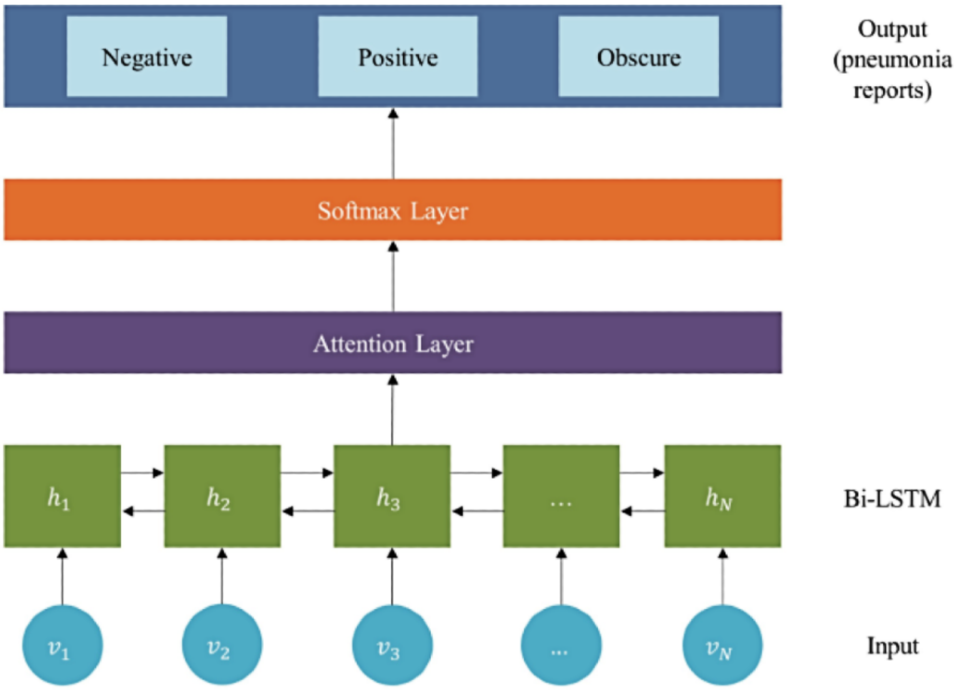
## Deep Learning–Based Classification Model

We built an attention-based deep neural network using LSTM. LSTM is a recurrent neural network variant that alleviates the vanishing gradient problem by learning and remembering long-term dependencies [18] and consists of a cell memory state and 3 gates.

The Bi-LSTM consists of a forward–backward LSTM layer [19]. Both layers are connected to the same output layer. Our classification model used Bi-LSTM with the attention mechanism. This allowed the model to simultaneously handle information from different positions.

Figure 2 shows the architecture of the deep learning–based classification model. First, the input is fed into the Bi-LSTM layer. Second, the output of the Bi-LSTM layer is fed into the attention layer (Bi-LSTM–Attention) for attending important words. Finally, the output of the attention weight passes through the softmax layer for classification.

**Figure 2.** The architectures of a deep learning-based classification model. Each input receives an embedding of English translated from Korean. In the attention layer, each word has an attention weight which is translated into the importance for prediction. Bi-LSTM: bidirectional long short-term memory model.



The performance metrics (ie, precision, recall [sensitivity], and $F_1$ score) were used to evaluate the models. The accuracy, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) were used to compare the models. For analyzing the multilabel data set, labels were treated as interested labels and other labels in evaluating each metric. For example, when we treated the precision for negative labels, only the true negative data were treated as true labels while positive and obscure labels were treated as false labels. $F_1$ score is the weighted average of precision and recall, and it is used to measure the performance of a model when the data consist of uneven class distributions [20]. The statistical analysis was performed on Python 3.7.6 (Python Software Foundation).

# Results

In this section, we evaluated the performance of the various classification models. To demonstrate the performance of our method, we compare the proposed model with traditional machine learning and other deep learning models. The machine learning models included logistic regression [21], support vector machine [22], Naïve Bayes regression [23], K-nearest neighbors algorithm [24], decision tree [25], and random forest [26]. The deep learning models included the word-to-vector representation model (Word2Vec) [27], FastText [17], CNN [28], and LSTM [29]. The details of each model are described in Multimedia Appendix 1.

Out of 5450 data sets, 4005 did not contain pneumonia, 895 contained pneumonia, and 550 were obscure, with 80% being used in the training set and the remaining 20% in test set. The test set was composed of no pneumonia (n=801), pneumonia (n=179), and obscure (n=110) classifications. The extra-validation set was annotated by 2 independent clinicians. Out of a total of 1000 radiologic reports, 803 labels were agreed upon by 2 independent clinicians. Among these labels, 498 did not contain pneumonia, 185 contained pneumonia, and 120 were obscure cases.

## Accuracy of Our Model as Compared to Previous Models

We evaluated the performance of the different models to find the best model. As shown in Table 1, the prediction accuracy changed depending on the model. The traditional models (ie, support vector machine, Naïve Bayes, etc) achieved an accuracy between 64.03% and 83.03%. The logistic regression showed a reasonable performance with an accuracy of 83.03% (Multimedia Appendix 1 Table S1).

The deep learning–based methods (ie, FastText, Word2Vec with Bi-LSTM–Attention, and the proposed model) outperformed the traditional models. The prediction accuracy of the deep learning models was 90.00%, 88.99%, and 91.01% for FastText, Word2Vec with Bi-LSTMAttention, and the proposed model, respectively. These deep learning models showed a 10% higher accuracy than did the traditional machine learning methods because sentence classification required the interpretation of complex features. The proposed model achieved the highest performance compared to the other deep learning models (Multimedia Appendix 1 Table S1).

## Model Accuracy Based on the Different Representation Methods of Words

We evaluated the performance based on different methods of word representation. The Word2Vec with Bi-LSTM–Attention model is a more commonly used language representation model. The model showed a higher accuracy and $F_1$ score than did the traditional models; however, the drawback associated with this model is that the foreign language is not represented (Table 1). We implemented another representation method with a substring using the FastText model. This method involves slicing of words to bunches of characters, which can be a better expression for the foreign language. The substring with FastText model achieved a precision of 93% for negative, 84% for positive, and 74% for obscure classifications; and a recall of 93% for negative, 84% for positive, and 47% for obscure classifications. The substring with FastText model showed a better performance than did the Word2Vec model according to $F_1$ score.

Our proposed model (Kor2Eng) translated Korean to English before the prediction process. The proposed model achieved a precision of 96%, 86%, and 61%, and a recall of 97%, 80%, and 64% for positive, negative, and obscure classifications, respectively. The AUROC of the model was 0.98 for negative, 0.97 for positive, and 0.90 for obscure classifications, while the AUPRC was 0.99 for negative, 0.87 for positive, and 0.62 for obscure classifications (Multimedia Appendix 1 Figure S5). Compared to the classification of the negative labels, which was a relatively easy task (96% of negative), classifying positive or obscure labels was a harder task and showed a rather lower $F_1$ score (83% for positive and 62% for obscure). For classifying the obscure classification, our model showed the highest performance among different representation methods (substring with FastText, Word2Vec, and Kor2Eng).

**Table 1.** The detailed performance of the top 3 best-performing models.

| Models | Precision, n/N (%) | Recall, n/N (%) | $F_1$ score (%) | AUROC[a] | AUPRC[b] |
|---|---|---|---|---|---|
| **Substring+FastText** [17] | | | | | |
| Negative | 776/819 (94.7) | 776/801 (96.9) | 96 | 0.82 | 0.92 |
| Positive | 153/593 (25.8) | 153/179 (85.5) | 83 | 0.74 | 0.34 |
| Obscure | 52/73 (71.2) | 52/110 (47.3) | 57 | 0.71 | 0.22 |
| **Word2Vec[c]+Bi-LSTM[d]–Attention** | | | | | |
| Negative | 772/849 (90.9) | 772/801 (96.4) | 94 | 0.95 | 0.98 |
| Positive | 153/222 (68.9) | 153/179 (85.5) | 81 | 0.96 | 0.87 |
| Obscure | 47/80 (58.8) | 47/110 (42.7) | 49 | 0.88 | 0.51 |
| **Proposed model (Kor2Eng[e])** | | | | | |
| Negative | 776/809 (95.9) | 776/801 (96.9) | 96 | 0.98 | 0.99 |
| Positive | 153/182 (84.1) | 153/179 (85.5) | 83 | 0.97 | 0.87 |
| Obscure | 70/115 (60.9) | 70/110 (63.6) | 62 | 0.90 | 0.62 |

[a] AUROC: area under the receiver operating characteristic curve.

[b] AUPRC: area under the precision-recall curve.

[c] Word2Vec: the word-to-vector representation model.

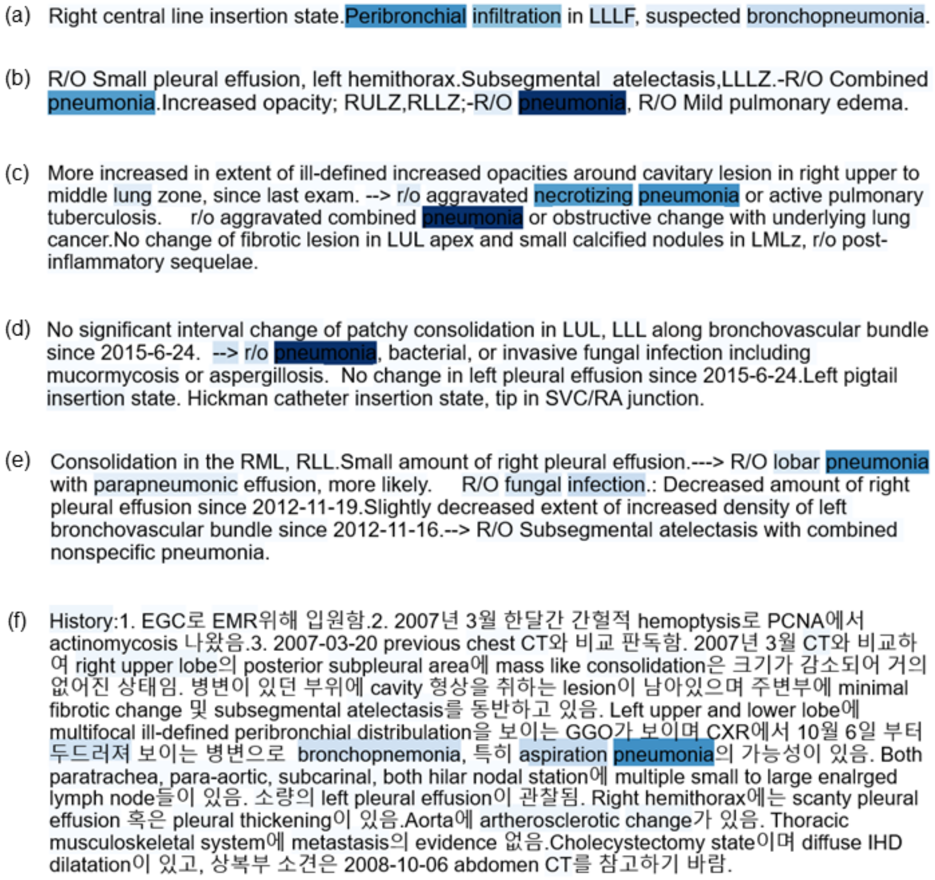
[d] Bi-LSTM: bidirectional long short-term memory model.

[e] Kor2Eng: Korean to English.

## Visualization of Relative Importance

We visualized the weighted words when the proposed model classified the input data. In the attention model, the weight of each word could be used for classifying the reports. Based on the intensity of color, the importance of a word was indicated when the proposed model determined the class of the input data. Darker colors indicated a higher importance for classifying pneumonia. Figure 3 shows the instances where the proposed model predicted pneumonia reports correctly. For example, the highlighted words “Peribronchial,” “infiltration,” “suspected,” and “bronchopneumonia” indicate pneumonia (Figure 3a). In the bilingual texts (Figure 3f), the following words are important to classifying pneumonia-reports: “두드러져,” “bronchopneumonia,” “aspiration,” and “pneumonia.”

**Figure 3.** Visualization of the importance of words by attention weights.The darker the color is, the greater the importance of the words for predicting the pneumonia label. High attention weight is depicted in the darker color. Words with high attention weights are shown.



(a) Right central line insertion state.Peribronchial infiltration in LLLF, suspected bronchopneumonia.

(b) R/O Small pleural effusion, left hemithorax.Subsegmental atelectasis,LLLZ.-R/O Combined pneumonia.Increased opacity; RULZ,RLLZ;-R/O pneumonia, R/O Mild pulmonary edema.

(c) More increased in extent of ill-defined increased opacities around cavitary lesion in right upper to middle lung zone, since last exam. --> r/o aggravated necrotizing pneumonia or active pulmonary tuberculosis. r/o aggravated combined pneumonia or obstructive change with underlying lung cancer.No change of fibrotic lesion in LUL apex and small calcified nodules in LMLz, r/o post-inflammatory sequelae.

(d) No significant interval change of patchy consolidation in LUL, LLL along bronchovascular bundle since 2015-6-24. --> r/o pneumonia, bacterial, or invasive fungal infection including mucormycosis or aspergillosis. No change in left pleural effusion since 2015-6-24.Left pigtail insertion state. Hickman catheter insertion state, tip in SVC/RA junction.

(e) Consolidation in the RML, RLL.Small amount of right pleural effusion.---> R/O lobar pneumonia with parapneumonic effusion, more likely. R/O fungal infection.: Decreased amount of right pleural effusion since 2012-11-19.Slightly decreased extent of increased density of left bronchovascular bundle since 2012-11-16.--> R/O Subsegmental atelectasis with combined nonspecific pneumonia.

(f) History:1. EGC로 EMR위해 입원함.2. 2007년 3월 한달간 간헐적 hemoptysis로 PCNA에서 actinomycosis 나왔음.3. 2007-03-20 previous chest CT와 비교 판독함. 2007년 3월 CT와 비교하여 right upper lobe의 posterior subpleural area에 mass like consolidation은 크기가 감소되어 거의 없어진 상태임. 병변이 있던 부위에 cavity 형상을 취하는 lesion이 남아있으며 주변부에 minimal fibrotic change 및 subsegmental atelectasis를 동반하고 있음. Left upper and lower lobe에 multifocal ill-defined peribronchial distribulation을 보이는 GGO가 보이며 CXR에서 10월 6일 부터 두드러져 보이는 병변으로 bronchopnemonia, 특히 aspiration pneumonia의 가능성이 있음. Both paratrachea, para-aortic, subcarinal, both hilar nodal station에 multiple small to large enalrged lymph node들이 있음. 소량의 left pleural effusion이 관찰됨. Right hemithorax에는 scanty pleural effusion 혹은 pleural thickening이 있음.Aorta에 artherosclerotic change가 있음. Thoracic musculoskeletal system에 metastasis의 evidence 없음.Cholecystectomy state이며 diffuse IHD dilatation이 있고, 상복부 소견은 2008-10-06 abdomen CT를 참고하기 바람.

## Extra Validations

As an extra validation of our proposed model, 2 clinicians labeled an additional data set. The data set was randomly selected from the entire data set, excluding the previously trained data. For precise labeling, 2 medical doctors each labeled the records. Of the 1000 records, 803 were agreed upon by 2 independent physicians. The Cohen κ value of the clinicians' label was 0.63 (95% CI 0.59-0.67). Table 2 shows the performance results of the proposed model with the extra-validation data set. The AUROC and AUPRC for positive labels were slightly lower in the extra-validation set than in the test set (Figure 4). The $F_1$ score of positive labels was similar to that of the training data; however, predicting negative and obscure labels showed a relatively poor performance as compared to the training data set according to $F_1$ score. The overall accuracy of our model was 80.0%.

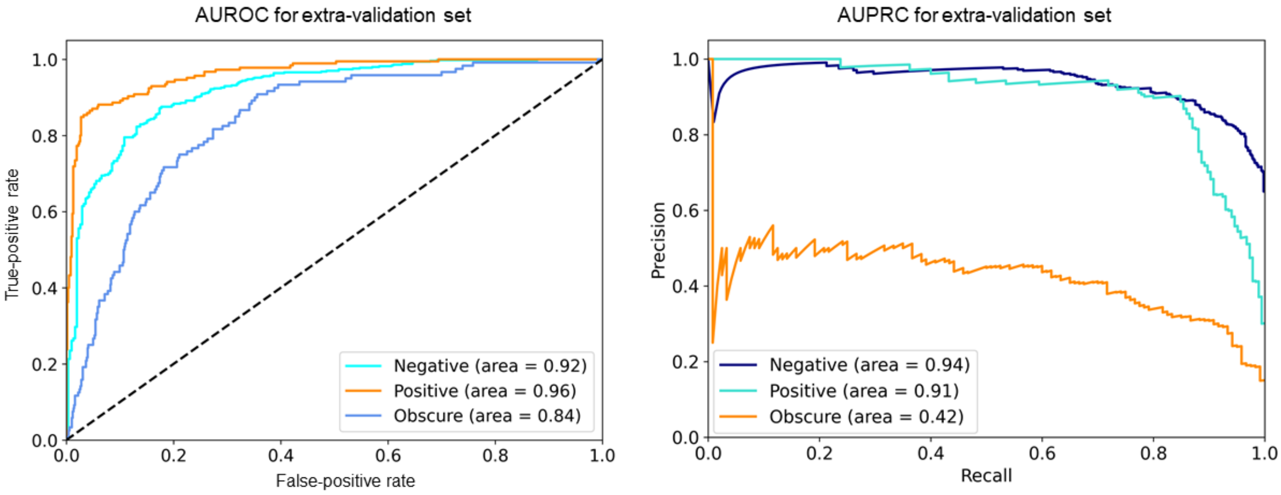
**Table 2.** Extra validation of the proposed Korean-to-English (Kor2Eng) model.

| Class | Precision, n/N (%) | Recall, n/N (%) | $F_1$ score | AUROC[a] | AUPRC[b] |
|---|---|---|---|---|---|
| Negative | 422/470 (89.8%) | 422/498 (84.7%) | 87% | 0.92 | 0.94 |
| Positive | 142/155 (91.6%) | 142/185 (76.8%) | 84% | 0.96 | 0.91 |
| Obscure | 77/178 (43.3%) | 77/120 (64.2%) | 52% | 0.84 | 0.42 |

[a] AUROC: area under the receiver operating characteristic curve.

[b] AUPRC: area under the precision-recall curve

**Figure 4.** AUROC and AUPRC of our proposed model in the extra-validation set. AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision-recall curve.



## Discussion

The purpose of the Kor2Eng model is to classify pneumonia-related medical records written in Korean and English. Our proposed model showed 91.01% accuracy in the test set and 80.0% accuracy in the extra-validation set for classifying pneumonia reports. Appropriate classification of radiologic reports is mandatory for further analysis regarding pneumonia through EMRs. As compared to other models, such as CNN or traditional machine learning models, our model showed better performance. The 3 best-performing models (Word2Vec with Bi-LSTM–Attention, FastText, and the proposed model) demonstrated better performance than did the traditional and CNN models, and our proposed model provided the highest AUROC and AUPRC among the top 3 models. Because too many false-positives may lead to clinician exhaustion, a model with excellent performance is desirable. We consider that a model with an AUROC of at least 0.95 can be used in clinical practice or for labeling the data set. The false-positive results of pneumonia reports can be additionally filtered with other clinical findings such as respiratory symptoms or antibiotics use, as pneumonia is defined by respiratory symptoms with radiologic findings [30].

The label balance of the data set was a consequence of excluding irrelevant labels to our target. As the reports that do not have pneumonia-related words can be considered pneumonia-negative radiologic reports, the reports requiring classification must contain at least one of the pneumonia-related words such as "consolidation" or "haziness". Excluding the irrelevant label is clinically appropriate and balances the data set with each label, with the balanced data set mitigating the overestimation of the model. Furthermore, filtering radiologic reports containing relevant words might make the data set rather homogenous, which makes classification a hard task. Our model showed an excellent performance in classifying pneumonia, and thus, it can be used for auto-labeling in classifying pneumonia reports.

A notable observation is the discrepancy between the test and extra-validation set. The model showed a rather similar performance in classifying negative and positive cases and a relatively poor performance in obscure cases. One reason for this discrepancy might be that 2 different clinicians annotated the entire extra-validation set. As some of the obscure cases are classified by the nuance of the context, the 2 clinicians might have differed in labeling the obscure cases. Therefore, the labeling of the obscure classification in the extra-validation set might have been different from that of the training set. The pneumonia cases in the report should only be decided by clinical situations, and thus, the importance of obscure cases should be evaluated in subsequent studies.

Several studies have been conducted for classifying radiologic reports as positive or negative for a given disease [1,10,31,32] or for classifying various diagnoses from medical records written in Chinese [13]. Most of the studies used a CNN-based model and showed a better performance than did our model [1,10,31,32]. In our study, we compared several deep learning models from logistic regression to LSTM with attention. The CNN model, which showed an excellent performance in previous studies [1,10,31,32], was inferior to the attention-based LSTM model in our data set. The reason for its relatively poor performance might be explained by our data selection. We selected radiologic reports that had at least one of the pneumonia-related words. This selection made the radiologic reports relatively homogeneous compared to those used in previous studies, which might contain a wider variety of radiologic reports. As we compared the performance with the CNN model, our proposed model was found to be comparably accurate with those of previous studies and showed better performance.

Radiologic reports in this study consisted of 2 languages: English and Korean. Compared to the English data set, the Korean word data set has a lack of studies in embedding and analyzing in deep learning. To overcome this limitation, we used unsupervised translation of Korean words to English words, which had pretrained embedding [17]. Compared to the Word2Vec with Bi-LSTM–Attention model, the attention/LSTM model with transfer embedding showed a better performance in classification, especially for obscure labels. This method might be especially important in bilingual reports.

Our study has several limitations. First, we only included reports from a single tertiary center of surgical in-patients. Our model might be inaccurate in a reporting style different from the one that we have incorporated. Thus, if the model used a data set from another reporting style, the model would need to be validated again. However, in this case, more labeled data might be available, and thus the applied method would show better performance in another data set, especially for bilingual text reports. Second, we could not compare the exact same models with the previous models that showed good performance. However, we compared our model with various deep learning models that were used in previous studies, which is sufficient to compare the performance of different model structures.

In summary, our proposed model showed superior performance as compared to other algorithms in the classification of pneumonia from radiologic reports. In bilingual radiologic reports, the proposed method of transferring and Bi-LSTM–Attention model showed significant improvement in performance than did the previous high-performing models. We hope that this method could be used to enrich the research about pneumonia by obtaining exact outcomes from electronic health data.

## Acknowledgments

This work was supported by the Bio-Synergy Research Project (no. NRF-2013M3A9C4078138) of the Ministry of Science, Information and Communications Technology, and Future Planning through the National Research Foundation.

## Authors' Contributions

HJP and CMC contributed to the conception and design of the study, as well as the data acquisition. HJP, BYS, EBL, and MS contributed to the analysis and interpretation of the data. HJP, BYS, EBL, and MS drafted the manuscript. HJP, CMC, and MS contributed to the critical revision of the paper, and all authors gave final approval for publication.

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Supplementary method and figures.
[DOCX File , 603 KB-Multimedia Appendix 1]

## References

1. Kehl KL, Elmarakeby H, Nishino M, Van Allen EM, Lepisto EM, Hassett MJ, et al. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. JAMA Oncol 2019 Oct 01;5(10):1421. [doi: 10.1001/jamaoncol.2019.1800]
2. Sherman RE, Anderson SA, Dal PGJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - What is it and what can it tell us? N Engl J Med 2016 Dec 08;375(23):2293-2297. [doi: 10.1056/NEJMsb1609216] [Medline: 27959688]
3. Park HJ, Jung DY, Ji W, Choi CM. Detection of bacteremia in surgical in-patients using recurrent neural network based on time series records: development and validation study. J Med Internet Res 2020 Aug 04;22(8):e19512 [FREE Full text] [doi: 10.2196/19512] [Medline: 32669261]
4. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med Inform 2016 Sep 30;4(3):e28 [FREE Full text] [doi: 10.2196/medinform.5909] [Medline: 27694098]
5. Saqib M, Sha Y, Wang MD. Early prediction of sepsis in EMR records using traditional ML techniques and deep learning LSTM networks. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul 04;2018(8):4038-4041 [FREE Full text] [doi: 10.1109/EMBC.2018.8513254] [Medline: 30441243]
6. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018 Apr 14;46(4):547-553 [FREE Full text] [doi: 10.1097/CCM.0000000000002936] [Medline: 29286945]
7. Huh K, Hong J, Jung J. Association of meteorological factors and atmospheric particulate matter with the incidence of pneumonia: an ecological study. Clin Microbiol Infect 2020 Dec;26(12):1676-1683. [doi: 10.1016/j.cmi.2020.03.006] [Medline: 32184173]
8. Liu WC, Lin CS, Yeh CC, Wu HY, Lee YJ, Chung CL, et al. Effect of influenza vaccination against postoperative pneumonia and mortality for geriatric patients receiving major surgery: a nationwide matched study. J Infect Dis 2018 Feb 14;217(5):816-826. [doi: 10.1093/infdis/jix616] [Medline: 29216345]
9. Mirończuk M. Information extraction system for transforming unstructured text data in fire reports into structured forms: a Polish case study. Fire Technol 2019 Jul 26;56(2):545-581. [doi: 10.1007/s10694-019-00891-z]

10. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, et al. Deep learning to classify radiology free-text reports. Radiology 2018 Mar;286(3):845-852. [doi: 10.1148/radiol.2017171115] [Medline: 29135365]
11. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. N Engl J Med 2016 Sep 29;375(13):1216-1219 [FREE Full text] [doi: 10.1056/NEJMp1606181] [Medline: 27682033]
12. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
13. Zhou S, Li X. Feature engineering vs. deep learning for paper section identification: toward applications in Chinese medical literature. Information Processing & Management 2020 May;57(3):102206. [doi: 10.1016/j.ipm.2020.102206]
14. Ong CJ, Orfanoudaki A, Zhang R, Caprasse FPM, Hutch M, Ma L, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. PLoS One 2020;15(6):e0234908 [FREE Full text] [doi: 10.1371/journal.pone.0234908] [Medline: 32559211]
15. Shin SY, Lyu Y, Shin Y, Choi HJ, Park J, Kim WS, et al. Lessons learned from development of de-identification system for biomedical research in a Korean tertiary hospital. Healthc Inform Res 2013 Jun;19(2):102-109 [FREE Full text] [doi: 10.4258/hir.2013.19.2.102] [Medline: 23882415]
16. Shin SY, Park YR, Shin Y, Choi HJ, Park J, Lyu Y, et al. A de-identification method for bilingual clinical texts of various note types. J Korean Med Sci 2015 Jan;30(1):7-15 [FREE Full text] [doi: 10.3346/jkms.2015.30.1.7] [Medline: 25552878]
17. Guillaume L, Alexis C, Ludovic D, Marc'Aurelio R. Unsupervised machine translation using monolingual corpora only. 2018 Presented at: Sixth International Conference on Learning Representations; Apr 30-May 3 2018; Vancouver URL: https://arxiv.org/abs/1711.00043
18. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
19. Dzmitry B, Kyunghyun C, Yoshua B. Neural machine translation by jointly learning to align and translate. 2015 Presented at: 3rd International Conference on Learning Representations; 2015 May 7-9; San Diego.
20. Lee SM, Seo JM, Yun J, Cho YH, Vogel-Claussen J, Schiebler ML, et al. Deep learning applications in chest radiography and computed tomography: current state of the art. J Thorac Imaging 2019 Mar;34(2):75-85. [doi: 10.1097/RTI.0000000000000387] [Medline: 30802231]
21. Cox D. The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) 2018 Dec 05;20(2):215-232. [doi: 10.1111/j.2517-6161.1958.tb00292.x]
22. Corinna C, Vladimir V. Support-vector networks. Machine learning 1995:273-297.
23. Sebastiani F. Machine learning in automated text categorization. ACM Comput. Surv 2002 Mar;34(1):1-47. [doi: 10.1145/505282.505283]
24. Soucy P, Mineau GW. A simple KNN algorithm for text categorization. : IEEE; 2001 Presented at: IEEE International Conference on Data Mining; Nov 29-Dec 2 2001; San Jose.
25. Quinlan J. Induction of decision trees. Mach Learn 1986 Mar;1(1):81-106. [doi: 10.1007/bf00116251]
26. Leo B. Random forests. Machine learning 2001 Jan:5-32.
27. Tomas M, Kai C, Greg C, Jeffrey D. Efficient estimation of word representations in vector space. 2013 Presented at: 2013 International Conference on Learning Representations,; May 2-May 4 2013; Scottsdale, AX.
28. Keiron O, Ryan N. An introduction to convolutional neural networks. 2015 Dec 2. URL: https://arxiv.org/abs/1511.08458 [accessed 2021-04-30]
29. Greff K, Srivastava RK, Koutnik J, Steunebrink BR, Schmidhuber K. LSTM: a search space odyssey. IEEE Trans Neural Netw Learn Syst 2017 Oct;28(10):2222-2232. [doi: 10.1109/TNNLS.2016.2582924] [Medline: 27411231]
30. Ranzani OT, Prina E, Menéndez R, Ceccato A, Cilloniz C, Méndez R, et al. New sepsis definition (sepsis-3) and community-acquired pneumonia mortality. A validation and clinical decision-making study. Am J Respir Crit Care Med 2017 Nov 15;196(10):1287-1297. [doi: 10.1164/rccm.201611-2262OC] [Medline: 28613918]
31. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, et al. A clinical text classification paradigm using weak supervision and deep representation. BMC Med Inform Decis Mak 2019 Jan 07;19(1):1 [FREE Full text] [doi: 10.1186/s12911-018-0723-6] [Medline: 30616584]
32. Shi X, Hu Y, Zhang Y, Li W, Hao Y, Alelaiwi A, et al. Multiple disease risk assessment with uniform model based on medical clinical notes. IEEE Access 2016;4:7074-7083. [doi: 10.1109/ACCESS.2016.2614541]

## Abbreviations

**AUPRC:** area under the precision-recall curve
**AUROC:** area under the receiver operating characteristic curve
**Bi-LSTM:** bidirectional long short-term memory model
**CNN:** convolutional neural network
**CT:** computed tomography
**EHR:** electronic health record

**ICD-9-CM:** International Classification of Diseases, Ninth Revision, Clinical Modification
**Kor2Eng:** Korean to English
**LSTM:** long short-term memory model
**Word2Vec:** word-to-vector representation model

*Edited by G Eysenbach; submitted 06.10.20; peer-reviewed by S Sheikhalishahi, M Aczon, R Abeysinghe, X Wang; comments to author 28.10.20; revised version received 21.12.20; accepted 04.04.21; published 17.05.21*

*Please cite as:*
*Park H, Song M, Lee EB, Seo BK, Choi CM*
*An Attention Model With Transfer Embeddings to Classify Pneumonia-Related Bilingual Imaging Reports: Algorithm Development and Validation*
*JMIR Med Inform 2021;9(5):e24803*
*URL: https://medinform.jmir.org/2021/5/e24803*
*doi: 10.2196/24803*
*PMID: 33820755*