Identification of Patients with Acute Lung Injury from Free-Text Chest X-Ray Reports

Meliha Yetisgen-Yildiz
University of Washington
Seattle, WA 98195
melihay@uw.edu

Cosmin Adrian Bejan
University of Washington
Seattle, WA 98195
bejan@uw.edu

Mark M. Wurfel
University of Washington
Seattle, WA 98195
mwurfel@uw.edu

Abstract

Identification of complex clinical phenotypes among critically ill patients is a major challenge in clinical research. The overall research goal of our work is to develop automated approaches that accurately identify critical illness phenotypes to prevent the resource intensive manual abstraction approach. In this paper, we describe a text processing method that uses Natural Language Processing (NLP) and supervised text classification methods to identify patients who are positive for Acute Lung Injury (ALI) based on the information available in free-text chest x-ray reports. To increase the classification performance we enhanced the baseline unigram representation with bigram and trigram features, enriched the n-gram features with assertion analysis, and applied statistical feature selection. We used 10-fold cross validation for evaluation and our best performing classifier achieved 81.70% precision (positive predictive value), 75.59% recall (sensitivity), 78.53% f-score, 74.61% negative predictive value, 76.80% specificity in identifying patients with ALI.

1 Introduction

Acute lung injury (ALI) is a critical illness consisting of acute hypoxemic respiratory failure with bilateral pulmonary infiltrates that is associated with pulmonary and non-pulmonary risk factors. ALI and its more severe form, acute respiratory distress syndrome (ARDS), represent a major health problem with an estimated prevalence of 7% of intensive care unit admissions (Rubenfeld et al., 2005) for which the appropriate treatment is often instituted too late or not at all (Ferguson et al., 2005; Rubenfeld et al., 2004). Early detection of ALI syndrome is essential for appropriate application of the only therapeutic intervention demonstrated to improve mortality in ALI, lung protective ventilation (LPV).

The identification of ALI requires recognition of a precipitating cause, either due to direct lung injury from trauma or pneumonia or secondary to another insult such as sepsis, transfusion, or pancreatitis. The consensus criteria for ALI include the presence of bilateral pulmonary infiltrates on chest radiograph, representing non-cardiac pulmonary edema as evidenced by the absence of left atrial hypertension (Pulmonary Capillary Wedge Pressure < 18 mmHg (2.4 kPa)) or absence of clinical evidence of congestive heart failure, and oxygenation impairment as defined by an arterial vs. inspired oxygen level ratio (PaO2/FiO2) <300 mmHg (40 kPa)) (Argitas et al., 1998; Dushianthan et al., 2011; Ranieri et al., 2012).

In this paper, we describe a text processing approach to identify patients who are positive for ALI based only on the free-text chest x-ray reports.

2 Related Work

Several studies demonstrated the value of Natural Language Processing (NLP) in a variety of health care applications including phenotype extraction from electronic medical records (EMR) (Demner-Dushman et al., 2009). Within this domain, chest x-ray reports have been widely studied to extract different types of pneumonia (Tepper et al., 2013; Elkin et al., 2008; Aronsky et al., 2001; Fiszman et al., 2000). Chest x-ray reports have also been studied for ALI surveillance by other researchers. Two of the prior studies relied on rule-based keyword search approaches. Herasevich et al. (2009) included a free text Boolean query containing trigger words *bilateral*, *infiltrate*, and *edema*. Azzam et al. (2009) used a more extensive list of trigger words and phrases to identify the presence of bilateral infiltrates and

ALI. In another study, Solti et al. (2009) represented the content of chest x-ray reports using character n-grams and applied supervised classification to identify chest x-ray reports consistent with ALI. In our work, different from prior research, we proposed a fully statistical approach where (1) the content of chest x-ray reports was represented by token n-grams, (2) statistical feature selection was applied to select the most informative features, and (3) assertion analysis was used to enrich the n-gram features. We also implemented Azzam et al.'s approach based on the information available in their paper and used it as a baseline to compare performance results of our approach to theirs.

3 Methods

The overall architecture of our text processing approach for ALI identification is illustrated in Figure 1. In the following sections, we will describe the main steps of the text processing approach as well as the annotated chest x-ray corpus used in training and test.

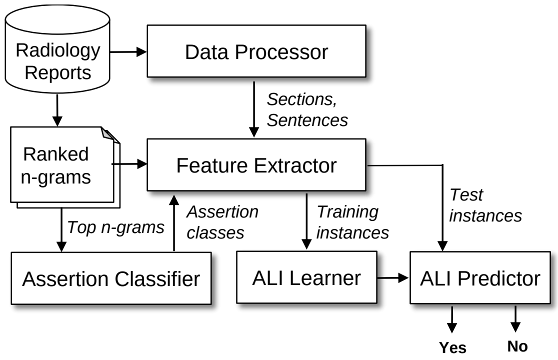

Figure 1 Overall system architecture of ALI extractor.

3.1 Chest X-ray Corpora

To develop the ALI extractor, we created a corpus composed of 1748 chest x-ray reports generated for 629 patients (avg number of reports=2.78, min=1, max=3). Subjects for this corpus were derived from a cohort of intensive care unit (ICU) patients at Harborview Medical Center that has been described previously (Glavan et al., 2011). We selected 629 subjects who met the oxygenation criteria for ALI (**PaO2/FiO2<300 mmHg**) and then three consecutive chest radiographs were pulled from the radiology database. Three Critical Care Medicine specialists reviewed the chest radiograph images for each patient and annotated the radiographs as consistent (positive) or not-consistent (negative) with ALI. We assigned ALI status for each subject based on the number of physician raters calling the chest radiographs consistent or not consistent with ALI. Table 1 shows the number of physicians with agreement on the radiograph interpretation. There were 254 patients in the positive set (2 or more physicians agreeing on ALI positive) and 375 patients in the negative set (2 or more physicians agreeing on ALI negative). Table 1 includes the distribution of patients over the positive and negative classes at different agreement levels. We will refer to this annotated corpus as the development set in the remaining of the paper.

Annotation	Agreement	Patient Count
ALI positive patients	3 2	147 107
ALI negative patients	3 2	205 170

Table 1 Agreement levels

For validation, we used a second dataset generated in a similar fashion to the development set. We obtained chest radiographs for 55 subjects that were admitted to ICU and who met oxygenation criteria for ALI (1 radiograph and report per patient). A specialized chest radiologist annotated each report for the presence of ALI. There were 21 patients in the positive set and 34 in the negative set. We will refer to this corpus as the validation set in the remaining of the paper.

The retrospective review of the reports in both corpora was approved by the University of Washington Human Subjects Committee of Institutional Review Board who waived the need for informed consent.

3.2 Pre-processing – Section and Sentence Segmentation

Although radiology reports are in free text format, they are somewhat structured in terms of sections. We used a statistical section segmentation approach we previously built to identify the boundaries of the sections and their types in our corpus of chest x-ray reports (Tepper et al., 2012). The section segmenter was trained and tested with a corpus of 100 annotated radiology reports and produced 93% precision, 91% recall and 92% f-score (5-fold cross validation).

After identifying the report sections, we used the OpenNLP [1] sentence chunker to identify the boundaries of sentences in the section bodies.

This pre-processing step identified 8,659 sections and 15,890 sentences in 1,748 reports of the development set and 206 sections and 414 sentences in 55 reports of the validation set. We used the section information to filter out the sections with clinician signatures (e.g., *Interpreted By*, *Contributing Physicians*, *Signed By*). We used the sentences to extract the assertion values associated with n-gram features as will be explained in a later section.

3.3 Feature Selection

Representing the information available in the free-text chest x-ray reports as features is critical in identifying patients with ALI. In our representation, we created one feature vector for each patient. We used unigrams as the baseline representation. In addition, we used bigrams and trigrams as features. We observed that the chest x-ray reports in our corpus are short and not rich in terms of medical vocabulary usage. Based on this observation, we decided not to include any medical knowledge-based features such as UMLS concepts or semantic types. Table 2 summarizes the number of distinct features for each feature type used to represent the 1,748 radiology reports for 629 patients.

Feature Type	# of Distinct Features
Unigram (baseline)	1,926
Bigram	10,190
Trigram	17,798

Table 2 Feature set sizes of the development set.

As can be seen from the table, for bigrams and trigrams, the feature set sizes is quite high. Feature selection algorithms have been successfully applied in text classification in order to improve the classification accuracy (Wenqian et al., 2007). In previous work, we applied statistical feature selection to the problem of pneumonia detection from ICU reports (Bejan et al., 2012). By significantly reducing the dimensionality of the feature space, they improved the efficiency of the pneumonia classifiers and provided a better understanding of the data.

We used statistical hypothesis testing to determine whether there is an association between a given feature and the two categories of our problem (i.e, positive and negative ALI). Specifically, we computed the χ^2 statistics (Manning and Schutze, 1999) which generated an ordering of features in the training set. We used 10-fold cross validation (development set) in our overall performance evaluation. Table 3 lists the top 15 unigrams, bigrams, and trigrams ranked by χ^2 statistics in one of ten training sets we used in evaluation. As can be observed from the table, many of the features are closely linked to ALI.

Unigram	Bigram	Trigram
Diffuse	diffuse lung	opacities consistent with
Atelectasis	lung opacities	diffuse lung opacities
Pulmonary	pulmonary edema	change in diffuse
Consistent	consistent with	lung opacities consistent
Edema	opacities consistent	in diffuse lung
Alveolar	in diffuse	with pulmonary edema
Opacities	diffuse bilateral	consistent with pulmonary
Damage	with pulmonary	low lung volumes
Worsening	alveolar damage	or alveolar damage
Disease	edema or	pulmonary edema pneumonia
Bilateral	low lung	diffuse lung disease
Clear	edema pneumonia	edema pneumonia no
Severe	or alveolar	diffuse bilateral opacities
Injury	lung disease	lungs are clear
Bibasilar	pulmonary opacities	lung volumes with

Table 3 Top 15 most informative unigrams, bigrams, and trigrams for ALI classification according to χ2 statistics.

Once the features were ranked and their corresponding threshold values (*N*) were established, we built a feature vector for each patient. Specifically, given the subset of *N* relevant features extracted from the ranked list of features, we considered in the representation of a given patient's feature vector only the features from the subset of relevant features that were also found in the chest x-ray reports of the patient. Therefore, the size of the feature space is equal to the size of relevant features subset (*N*) whereas the length of each feature vector will be at most this value.

3.4 Assertion Analysis

We extended our n-gram representation with assertion analysis. We built an assertion classifier (Bejan et al., 2013) based on the annotated corpus of 2010 Integrating Biology and the Beside (i2b2) / Veteran's Affairs (VA) NLP challenge (Uzuner et al., 2011). The 2010 i2b2/VA challenge introduced assertion classification as a

[1] OpenNLP. Available at: http://opennlp.apache.org/

shared task, formulated such that each medical concept mentioned in a clinical report (e.g., asthma) is associated with a specific assertion category (present, absent, conditional, hypothetical, possible, and not associated with the patient). We defined a set of novel features that uses the syntactic information encoded in dependency trees in relation to special cue words for these categories. We also defined features to capture the semantics of the assertion keywords found in the corpus and trained an SVM multi-class classifier with default parameter settings. Our assertion classifier outperformed the state-of-the-art results and achieved 79.96% macro-averaged F-measure and 94.23% micro-averaged F-measure on the i2b2/VA challenge test data.

For each n-gram feature (e.g., *pneumonia*), we used the assertion classifier to determine whether it is present or absent based on contextual information available in the sentence the feature appeared in (e.g., Feature: pneumonia, Sentence: There is no evidence of *pneumonia*, congestive heart failure, or other acute process., Assertion: absent). We added the identified assertion value to the feature (e.g., *pneumonia_absent*). The frequencies of each assertion type in our corpus are presented in Table 4. Because chest x-rays do not include family history, there were no instances of not associated with the patient. We treated the three assertion categories that express hedging (conditional, hypothetical, possible) as the present category.

Assertion Class	Frequency
Present	206,863
Absent	13,961
Conditional	4
Hypothetical	330
Possible	3,980

Table 4 Assertion class frequencies.

3.5 Classification

For our task of classifying ALI patients, we picked the Maximum Entropy (MaxEnt) algorithm due to its good performance in text classification tasks (Berger et al., 1996). In our experiments, we used the MaxEnt implementation in a machine learning package called Mallet[2].

4 Results

4.1 Metrics

We evaluated the performance by using precision (positive predictive value), recall (sensitivity), negative predictive value, specificity, f-score, and accuracy. We used 10-fold cross validation to measure the performance of our classifiers on the development set. We evaluated the best performing classifier on the validation set.

4.2 Experiments with Development Set

We designed three groups of experiments to explore the effects of (1) different n-gram features, (2) feature selection, (3) assertion analysis of features on the classification of ALI patients. We defined two baselines to compare the performance of our approaches. In the first baseline, we implemented the Azzam et. al.'s rule-based approach (2009). In the second baseline, we only represented the content of chest x-ray reports with unigrams.

4.3 N-gram Experiments

Table 5 summarizes the performance of n-gram features. When compared to the baseline *unigram* representation, gradually adding bigrams (*uni+bigram*) and trigrams (*uni+bi+trigram*) to the baseline increased the precision and specificity by 4%. Recall and NPV remained the same. Azzam et. al.'s rule-based baseline generated higher recall but lower precision when compared to n-gram features. The best f-score (64.45%) was achieved with the *uni+bi+trigram* representation.

4.4 Feature Selection Experiments

To understand the effect of large feature space on classification performance, we studied how the performance of our system evolves for various threshold values (*N*) on the different combinations of χ^2 ranked unigram, bigram, and trigram features. Table 6 includes a subset of the results we collected for different values of *N*. As listed

System configuration	TP	TN	FP	FN	Precision/ PPV	Recall/ Sensitivity	NPV	Specificity	F-Score	Accuracy
Baseline#1–Azzam et. al. (2009)	201	184	191	53	51.27	79.13	77.64	49.07	62.23	61.21
Baseline#2–*unigram*	156	288	87	98	64.20	61.42	74.61	76.80	62.78	70.59
Uni+bigram	156	296	79	98	66.38	61.42	75.13	78.93	63.80	71.86
Uni+bi+trigram	155	303	72	99	68.28	61.02	75.37	80.80	64.45	72.81

Table 5 Performance evaluation on development set with no feature selection. TP: True positive, TN: True negative, FP: False positive, FN: False negative, PPV: Positive predictive value, NPV: Negative predictive value. The row with the heighted F-Score is highlighted.

in this table, for *N*=100, the *unigram* representation performed better than *uni+bigram*, *uni+bi+trigram* feature combinations; however, as *N* increased, the performance of *uni+bi+trigram* performed better, reaching the best f-score (78.53%) at *N*=800. When compared to the two defined baselines, the performance results of *uni+bi+trigram* at *N*=800 were significantly better than those of the baselines.

4.5 Assertion Analysis Experiments

We ran a series of experiments to understand the effect of assertion analysis on the classification performance. We used the best performing clas-

N	Feature configuration	TP	TN	FP	FN	Precision/ PPV	Recall/ Sensitivity	NPV	Specificity	F-Score	Accuracy
100	Unigram	191	316	59	63	76.40	75.20	83.38	84.27	75.79	80.60
	Uni+bigram	180	313	62	74	74.38	70.87	80.88	83.47	72.58	78.38
	Uni+bi+trigram	183	317	58	71	75.93	72.05	81.70	84.53	73.94	79.49
200	Unigram	189	312	63	65	75.00	74.41	82.76	83.20	74.70	79.65
	Uni+bigram	183	321	54	71	77.22	72.05	81.89	85.60	74.54	80.13
	Uni+bi+trigram	190	322	53	64	78.19	74.80	83.42	85.87	76.46	81.40
300	Unigram	185	311	64	69	74.30	72.83	81.84	82.93	73.56	78.86
	Uni+bigram	188	322	53	66	78.01	74.02	82.99	85.87	75.96	81.08
	Uni+bi+trigram	187	331	44	67	80.95	73.62	83.17	88.27	77.11	82.35
400	Unigram	179	315	60	75	74.90	70.47	80.77	84.00	72.62	78.54
	Uni+bigram	184	319	56	70	76.67	72.44	82.01	85.07	74.49	79.97
	Uni+bi+trigram	184	325	50	70	78.63	72.44	82.28	86.67	75.41	80.92
500	Unigram	177	310	65	77	73.14	69.69	80.10	82.67	71.37	77.42
	Uni+bigram	178	321	54	76	76.72	70.08	80.86	85.60	73.25	79.33
	Uni+bi+trigram	187	325	50	67	78.90	73.62	82.91	86.67	76.17	81.40
600	Unigram	179	305	70	75	71.89	70.47	80.26	81.33	71.17	76.95
	Uni+bigram	177	320	55	77	76.29	69.69	80.60	85.33	72.84	79.01
	Uni+bi+trigram	189	325	50	65	79.08	74.41	83.33	86.67	76.67	81.72
700	Unigram	176	308	67	78	72.43	69.29	79.79	82.13	70.82	76.95
	Uni+bigram	180	323	52	74	77.59	70.87	81.36	86.13	74.07	79.97
	Uni+bi+trigram	189	328	47	65	80.08	74.41	83.46	87.47	77.14	82.19
800	Unigram	172	311	64	82	72.88	67.72	79.13	82.93	70.20	76.79
	Uni+bigram	180	327	48	74	78.95	70.87	81.55	87.20	74.69	80.60
	Uni+bi+trigram	192	332	43	62	81.70	75.59	84.26	88.53	78.53	83.31
900	Unigram	174	311	64	80	73.11	68.50	79.54	82.93	70.73	77.11
	Uni+bigram	182	328	47	72	79.48	71.65	82.00	87.47	75.36	81.08
	Uni+bi+trigram	187	333	42	67	81.66	73.62	83.25	88.80	77.43	82.67
1000	Unigram	177	313	62	77	74.06	69.69	80.26	83.47	71.81	77.90
	Uni+bigram	185	326	49	69	79.06	72.83	82.53	86.93	75.82	81.24
	Uni+bi+trigram	190	327	48	64	79.83	74.80	83.63	87.20	77.24	82.19

Table 6 Performance evaluation on development set with feature selection. TP: True positive, TN: True negative, FP: False positive, FN: False negative, PPV: Positive predictive value, NPV: Negative predictive value. The row with the heighted F-Score is highlighted.

Assertion configuration	TP	TN	FP	FN	Precision/ PPV	Recall/ Sensitivity	NPV	Specificity	F-Score	Accuracy
Assertion_none	192	332	43	62	81.70	75.59	84.26	88.53	78.53	83.31
Assertion_all	188	328	47	66	80.00	74.02	83.25	87.47	76.89	82.03
Assertion_top_10	191	328	47	63	80.25	75.20	83.89	87.47	77.64	82.51
Assertion_top_20	190	329	46	64	80.51	74.80	83.72	87.73	77.55	82.51
Assertion_top_30	190	331	44	64	81.20	74.80	83.80	88.27	77.87	82.83
Assertion_top_40	190	328	47	64	80.17	74.80	83.67	87.47	77.39	82.35
Assertion_top_50	190	330	45	65	80.85	74.51	83.54	88.00	77.55	82.54

Table 7 Performance evaluation on development set with the assertion feature (uni+bi+trigram at *N*=800). TP: True positive, TN: True negative, FP: False positive, FN: False negative, PPV: Positive predictive value, NPV: Negative predictive value. The row with the heighted F-Score is highlighted.

System configuration	TP	TN	FP	FN	Precision/ PPV	Recall/ Sensitivity	NPV	Specificity	F-Score	Accuracy
Baseline#1–Azzam et. al. (2009)	10	18	16	11	38.46	47.62	62.07	52.94	42.55	50.91
Baseline#2–*unigram*	12	29	5	9	70.53	57.14	76.32	85.29	63.16	74.55
Uni+bi+trigram at k=800	9	30	4	12	69.23	42.86	71.43	88.24	52.94	70.91

Table 8 Performance evaluation on validation set. TP: True positive, TN: True negative, FP: False positive, FN: False negative, PPV: Positive predictive value, NPV: Negative predictive value. The row with the heighted F-Score is highlighted.

sifier with *uni+bi+trigram* at *N*=800 in our experiments. We applied assertion analysis to all 800 features as well as only a small set of top ranked 10×k (1≤k≤5) features which were observed to be closely related to ALI (e.g., *diffuse*, *opacities*, *pulmonary edema*). We hypothesized applying assertion analysis would inform the classifier on the presence and absence of those terms which would potentially decrease the false positive and negative counts.

Table 7 summarizes the results of our experiments. When we applied assertion analysis to all 800 features, the performance slightly dropped when compared to the performance with no assertion analysis. When assertion analysis applied to only top ranked features, the best f-score performance was achieved with assertion analysis with top 30 features; however, it was still slightly lower than the f-score with no assertion analysis. The differences are not statistically significant.

4.6 Experiments with Validation Set

We used the validation set to explore the generalizability of the proposed approach. To accomplish this we run the best performing classifier (*uni+bi+trigram* at *N=800*) and two defined baselines on the validation set. We re-trained the *uni+bi+trigram* at *N=800* classifier and unigram baseline on the complete development set.

Table 8 includes the performance results. The second baseline with unigrams performed the best and Azzam et. al.'s baseline performed the worst in identifying the patients with ALI in the validation set.

5 Discussion

Our best system achieved an f-score of 78.53 (precision=81.70, recall=75.59) on the development set. While the result is encouraging and significantly better than the f-score of a previously published system (f-score=62.23, precision=51.27, recall=79.13), there is still room for improvement.

There are several important limitations to our current development dataset. First, the annotators who are pulmonary care specialists used only the x-ray images to annotate the patients. However, the classifiers were trained based on the features extracted from the radiologists' free-text interpretation of the x-ray images. In one false positive case, the radiologist has written *"Bilateral diffuse opacities, consistent with pulmonary edema. Bibasilar atelectasis."* in the chest x-ray report, however all three pulmonary care specialists annotated the case as negative based on their interpretation of images. Because the report consisted of many very strong features indicative of ALI, our classifier falsely identified the patient as positive with a very high prediction probability 0.96. Second, although three annotators annotated the development set, there was full agreement on 42.12% (107/254) of the positive patients and 45.33% (170/375) of the negative patients. Table 9 includes the false positive and negative statistics of the best performing classifier (*uni+bi+trigrams* at *N=800*). As can be seen from the table, the classifier made more mistakes on patients where the annotator agreement was not perfect. The classifier predicted 13 of the 28 false positives and 23 of the 39 false negatives with probabilities higher than 0.75. When we investigated the reports of those 13 false positives, we observed that the radiologists used many very strong ALI indicative features (e.g., *diffuse lung opacities*, *low lung volumes*) to describe the images. On the contrary, radiologists did not use as many ALI indicative features in the reports of 23 false negative cases.

Error Type	Agreement	Frequency	Percentage
False Positives	3	15	10.20% (15/147)
	2	28	26.17% (28/107)
False Negatives	3	24	11.70% (24/205)
	2	39	22.94% (39/170)

Table 9 False positive and false negative statistics at different agreement levels.

In our experiments on the development set, we demonstrated the positive impact of statistical feature selection on the overall classification performance. We achieved the best f-score, when we used only 2.67% (800/29,914) of the complete n-gram feature space. We enriched the highly ranked features with assertion analysis. However, unlike feature selection, assertion analysis did not improve the overall performance. To explore the reasons, we analyzed reports from our corpus and found out that the current six assertion classes (*present, absent, conditional, hypothetical, possible*) were not sufficient to capture true meaning in many cases. For example, our assertion classifier assigned the class *present* to the bigram *bibasilar opacities* based on the sentence "*There are bibasilar opacities that are unchanged*". Although *present* was the correct assignment for *bibasilar opacities*, the more important piece of information was the change of state in *bibasilar opacities* for ALI diagnosis. X-rays describe a single snapshot of time but the x-ray report narrative makes explicit or, more often implicit references to a previous x-ray. In this way, the sequence of x-ray reports is used not only to assess a patient's health at a moment in time but also to monitor the change. We recently defined a schema to annotate change of state for clinical events in chest x-ray reports (Vanderwende et al., 2013). We will use this annotation schema to create an annotated corpus for training models to enrich the assertion features for ALI classification.

The results on the validation set revealed that the classification performance degraded significantly when training and test data do not come from the same dataset. There are multiple reasons to this effect. First, the two datasets had different language characteristics. Although both development and validation sets included chest x-ray reports, only 2,488 of the 3,305 (75.28%) n-gram features extracted from the validation set overlapped with the 29,914 n-gram features extracted from the development set. We suspect that this is the main reason why our best performing classifier with feature selection trained on the development set did not perform as well as the unigram baseline on the validation set. Second, the validation set included only 55 patients and each patient had only one chest x-ray report unlike the development set where each patient had 2.78 reports on the average. In other words, the classifiers trained on the development set with richer content made poor predictions on the validation set with more restricted content. Third, because the number of patients in the validation set was too small, each false positive and negative case had a huge impact on the overall performance.

6 Conclusion

In this paper, we described a text processing approach to identify patients with ALI from the information available in their corresponding free-text chest x-ray reports. To increase the classification performance, we (1) enhanced the baseline unigram representation with bigram and trigram features, (2) enriched the n-gram features with assertion analysis, and (3) applied statistical feature selection. Our proposed methodology of ranking all the features using statistical hypothesis testing and selecting only the most relevant ones for classification resulted in significantly improving the performance of a previous system for ALI identification. The best performing classifier achieved 81.70% precision (positive predictive value), 75.59% recall (sensitivity), 78.53% f-score, 74.61% negative predictive value, 76.80% specificity in identifying patients with ALI when using the uni+bi+trigram representation at N=800. Our experiments showed that assertion values did not improve the overall performance. For future work, we will work on defining new semantic features that will enhance the current assertion definition and capture the change of important events in radiology reports.

Acknowledgements

The work is partly supported by the Institute of Translational Health Sciences (UL1TR000423), and Microsoft Research Connections. We would also like to thank the anonymous reviewers for helpful comments.

References

Aronsky D, Fiszman M, Chapman WW, Haug PJ. Combining decision support methodologies to diagnose pneumonia. AMIA Annu Symp Proc., 2001:12-16.

Artigas A, Bernard GR, Carlet J, Dreyfuss D, Gattinoni L, Hudson L, Lamy M, Marini JJ, Matthay MA, Pinsky MR, Spragg R, Suter PM. The American-European Consensus Conference on ARDS, part 2: Ventilatory, pharmacologic, supportive therapy, study design strategies, and issues related to recovery and remodeling. Acute respiratory distress syndrome. Am J Respir Crit Care Med. 1998;157(4 Pt1):1332-47.

Azzam HC, Khalsa SS, Urbani R, Shah CV, Christie JD, Lanken PN, Fuchs BD. Validation study of an automated electronic acute lung injury screening tool. J Am Med Inform Assoc. 2009; 16(4):503-8.

Bejan CA, Xia F, Vanderwende L, Wurfel M, Yetisgen-Yildiz M. Pneumonia identification using statistical feature selection. J Am Med Inform Assoc. 2012; 19(5):817-23.

Bejan CA, Vanderwende L, Xia F, Yetisgen-Yildiz M. Assertion Modeling and its role in clinical phenotype identification. J Biomed Inform. 2013; 46(1):68-74.

Berger AL, Pietra SAD, Pietra VJD. A maximum entropy approach to natural language processing. Journal of Computational Linguistics. 1996; 22(1):39-71.

Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform. 2009; 42(5):760-72.

Dushianthan A, Grocott MPW, Postle AD, Cusack R. Acute respiratory distress syndrome and acute lung injury. Postgrad Med J. 2011; 87:612-622.

Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H, Asatryan AX, Tokars JI, Rosenbloom ST, Brown SH. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc. 2008; 6:172-6.

Ferguson ND, Frutos-Vivar F, Esteban A, Fernández-Segoviano P, Aramburu JA, Nájera L, Stewart TE. Acute respiratory distress syndrome: underrecognition by clinicians and diagnostic accuracy of three clinical definitions. Crit Care Med. 2005; 33(10):2228-34.

Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc. 2000;7(6):593-604.

Glavan BJ, Holden TD, Goss CH, Black RA, Neff MJ, Nathens AB, Martin TR, Wurfel MM; ARDSnet Investigators. Genetic variation in the FAS gene and associations with acute lung injury. Am J Respir Crit Care Med. 2011;183(3):356-63.

Herasevich V, Yilmaz M, Khan H, Hubmayr RD, Gajic O. Validation of an electronic surveillance system for acute lung injury. Intensive Care Med. 2009; 35(6):1018-23.

Manning CD, Schutze H. Foundations of statistical natural language processing. MIT Press 1999.

Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, Fan E, Camporota L, Slutsky AS. Acute Respiratory Distress Syndrome. The Berlin Definition. JAMA. 2012; 307(23): 2526-2533.

Rubenfeld GD, Caldwell E, Peabody E, Weaver J, Martin DP, Neff M, Stern EJ, Hudson LD. Incidence and outcomes of acute lung injury. N Engl J Med. 2005; 353(16):1685-93.

Rubenfeld GD, Cooper C, Carter G, Thompson BT, Hudson LD. Barriers to providing lung-protective ventilation to patients with acute lung injury. Crit Care Med. 2004; 32(6):1289-93.

Solti I, Cooke CR, Xia F, Wurfel MM. Automated Classification of Radiology Reports for Acute Lung Injury: Comparison of Keyword and Machine Learning Based Natural Language Processing Approaches. Proceedings (IEEE Int Conf Bioinformatics Biomed). 2009;314-319.

Tepper M, Capurro D, Xia F, Vanderwende L, Yetisgen-Yildiz M. Statistical Section Segmentation in Free-Text Clinical Records. Proceedings of the International Conference on Language Resources and Evaluation (LREC), Istanbul, May 2012.

Tepper M, Evans HL, Xia F, Yetisgen-Yildiz M. Modeling Annotator Rationales with Application to Pneumonia Classification. Proceedings of Expanding the Boundaries of Health Informatics Using AI Workshop of AAAI'2013, Bellevue, WA; 2013.

Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011; 18(5):552–556.

Vanderwende L, Xia F, Yetisgen-Yildiz M. Annotating Change of State for Clinical Events. Proceedings of the 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation Workshop of NAACL'2013, Atlanta, June 2013.

Wenqian W, Houkuan H, Haibin Z et al. A novel feature selection algorithm for text categorization. Expert Syst Appl 2007;33:1–5.