Automated Classification of Radiology Reports for Acute Lung Injury: Comparison of Keyword and Machine Learning Based Natural Language Processing Approaches

1) Imre Solti, MD, PhD
Department of Medical Education and Biomedical Informatics, University of Washington, Seattle WA
solti@uw.edu

2) Colin R. Cooke, MD, MSc
Division of Pulmonary & Critical care Medicine, University of Michigan, Ann Arbor, MI
cookecr@umich.edu

3) Fei Xia, PhD
Department of Linguistics, University of Washington, Seattle, WA
fxia@uw.edu

4) Mark M. Wurfel, MD, PhD
Division of Pulmonary and Critical Care Medicine, University of Washington, Seattle, WA
mwurfel@uw.edu

Abstract

This paper compares the performance of keyword and machine learning-based chest x-ray report classification for Acute Lung Injury (ALI). ALI mortality is approximately 30 percent. High mortality is, in part, a consequence of delayed manual chest x-ray classification. An automated system could reduce the time to recognize ALI and lead to reductions in mortality. For our study, 96 and 857 chest x-ray reports in two corpora were labeled by domain experts for ALI. We developed a keyword and a Maximum Entropy-based classification system. Word unigram and character n-grams provided the features for the machine learning system. The Maximum Entropy algorithm with character 6-gram achieved the highest performance (Recall=0.91, Precision=0.90 and F-measure=0.91) on the 857-report corpus. This study has shown that for the classification of ALI chest x-ray reports, the machine learning approach is superior to the keyword based system and achieves comparable results to highest performing physician annotators.

1. Introduction

Acute Lung Injury (ALI) is a disease characterized by severe inflammation and fluid accumulation in the lungs leading to respiratory failure necessitating mechanical ventilation. Patients suffering from ALI have a mortality of approximately 30 percent and account for up to 75,000 deaths in the United States each year [1]. Accurately diagnosing ALI is complicated by the fact that it is a syndrome defined by multiple clinical features, including diffuse bilateral opacities on the patient's chest x-ray.

A radiologist and/or critical care physician interprets and classifies the chest x-ray report as consistent with ALI or not. The report's text itself very rarely mentions the patient's ALI status explicitly. Inconsistencies and delays in assessing the dictated reports can lead to delays in the management of patients with ALI and consequently to higher mortality.

As part of a larger ALI project, our aim is to design a reliable natural language processing (NLP) system to automate the classification of chest x-ray reports for ALI. In this paper we compare the performance of an NLP-based algorithm with a keyword-based classification system. In the next section, we describe the two corpora, the corresponding gold standards and the classification methods. In Section 3, we present our findings. In Section 4, we discuss the results. In Section 5, we describe the earlier published work in ALI classification. Finally, in Section 6, we provide assessment of our work and describe some of the limitations and future directions for the research.

2. Data and Methods

2.1. Corpus development

We selected two sets of files to create the corpora. First, we utilized an ongoing IRB-approved prospective collection of patients with respiratory failure and ALI at our institution. We selected a convenience sample of 52 patients with confirmed ALI diagnosis (cases) and a sample of 44 mechanically ventilated patients without ALI diagnosis (controls) from the same repository.

The second set included reports from 287 participants in an ongoing cohort study of patients with sepsis and hypoxemic respiratory failure. For each patient in the second corpus, multiple chest x-rays were selected. A total of 857 films were available for analysis in the second corpus.

Using the patients' medical records we identified the single chest x-ray (or multiple chest x-rays) temporally closest to the onset of hypoxemic respiratory failure. We then abstracted the dictated report for each chest x-ray including the "Impression" section. We processed the two corpora separately because of the differences in the patients from which the reports arose, the method of report selection (as described above), the number of reports per patient, and differences in the gold standard development for the two corpora (see below).

2.2. Development of the gold standard

The gold standard for the 96-report corpus was developed prior to report processing. Eleven pulmonary and critical care fellows and faculty independently read the textual report for each chest x-ray and classified it as consistent with ALI or not. The 11 annotators were blinded to the patient's true ALI diagnosis to assure the validity of the gold standard.

The patient's ALI diagnosis should not be used in a text classification because this information would not be available in a prospective diagnostic system. The final label of a report is determined by majority vote; that is, a report is classed as "yes" if at least 6 annotators say so.

For the 857-report corpus, one of the investigators (a critical care physician) generated the gold standard similarly to the process for the 96-report set. The annotator was blinded to the true ALI diagnosis of the patient.

2.3. Preprocessing steps

During preprocessing we removed date and time and adjusted the text for occasionally missing sentence boundary markings. The chest x-ray reports are dictated in our institution and the text is relatively well formed. Consequently we broke text into sentences by running a simple sentence boundary detector.

In the last preprocessing step we detected negations. For negation detection we relied on Chapman's NegEx algorithm [2]. We used Solti's GenNegEx Java implementation that achieves 98% accuracy on 2,376 sentences of the test kit [3]. Text that was within the negated scope of a sentence was removed. We used the default NegEx trigger terms of the kit.

2.4. Classification methods

We used two methods for document classification. In the first approach, a list of keywords was created manually by three domain experts (pulmonary care specialists). The experts worked independently and their lists were merged by one of the investigators. The final list included 48 phrases.

The experts expected that the presence of any of the keywords in a chest x-ray report increased the likelihood that the patient to whom the report belongs has ALI. Therefore, the keyword-based algorithm classifies a document as consistent with ALI if the number of occurrences of keywords in the document is no less than a threshold. We empirically set the threshold to be two.

Two of the experts also assigned weights on a scale of 1-3 and 1-10 to each term on their list. Higher weights denote a greater likelihood of ALI consistent text. We computed the average of the two experts' opinion for the final weight. Given the keyword list with associated weights, the score of a chest x-ray report is calculated with Equation 1, where W_i is the score for the i-th document, w_j is the weight of the j-th term, and $c_{i,j}$ is the occurrence of the j-th term in the i-th document.

Equation 1: $W_i = \Sigma(w_j * c_{i,j})$.

In the second approach, we apply standard classification algorithms to the task. The Maximum Entropy (MaxEnt) algorithm was selected because it allows overlapping features (i.e. features that are not conditionally independent from each other given the class labels) [4]. The other advantage of MaxEnt is that the feature weights estimated by MaxEnt could be useful to humans in understanding the importance of certain features.

For MaxEnt training and decoding we used the MALLET, a common machine learning package developed by the University of Massachusetts at Amherst [5]. Word unigrams and character n-grams were collected from each document as features. For statistical analysis we used SPSS 13.0 for Windows [6]. Equations 2-5 show the formulas for calculating recall, precision, harmonic F-measure and accuracy where TP denotes “True Positive” classification, FP indicates “False Positive”, FN indicates “False Negative” and TN denotes “True Negative” classification while P and R are precision and recall.

Equation 2: Recall = TP / (TP + FN)

Equation 3: Precision = TP / (TP + FP)

Equation 4: F-measure = 2 * P * R / (P +R)

Equation 5: Accuracy = (TP + TN) / (TP+TN+FP+FN)

For the MaxEnt classifier, we used 10-fold cross validation: we randomly assigned 90% of the corpus to training and 10% to test set. The results of the test data were averaged over the ten runs.

To get a better sense if the higher values for recall, precision and F-measure for some of the character n-grams were not purely chance based, we calculated the ROC curve for the larger corpus. We computed the area under the curve, p-values and the lower and upper bounds of the 95% confidence intervals for the word unigram, character n-gram and keyword and weight-based systems. To find where the accuracy for character n-gram based systems peaks we generated accuracy statistics for 1-gram to 14-gram. To find the peak we used the 857-report corpus.

2.5. Baseline

We calculated the baseline for the classification. We assigned ALI consistent class label to each chest x-ray report in the larger corpus as a default value without any processing. In this case the recall is 1.0 because obviously we will find all ALI consistent reports (392/392 = 1.0). The precision will be 0.46 (392/857 = 0.46). The F-measure for the baseline is 0.63 for the 857-report set. Similarly, the smaller corpus’ baseline has a recall of 1.0, a precision of 0.52 and the F-measure is 0.68.

3. Experimental Results

3.1. Descriptive statistics for the corpus

Table 1. shows the number of records, lines and words for the two corpora before and after preprocessing.

Table 1. Results of preprocessing the corpora

	96-report		**857-report**	
	pre	**post**	**pre**	**post**
Records	96	96	857	855
Lines	2,215	894	12,826	7,167
Words	6,756	4,870	46,537	35,772

During preprocessing we had to eliminate two records from the 857-report set because the two files included multiple collated reports. Preprocessing for removal of billing code, date, time and negation scope reduced the size of the first corpora by 28% and the second corpora by 23%.

Approximately 20% of the original text was eliminated as part of a negation scope to remove negated clinical findings. For example, a record might have included the following sentence: “Bilateral opacities no edema.” The negation detection phase of the preprocessing step removed “no edema” and our system processed the rest of the sentence: “Bilateral opacities”.

3.2. Gold standard

The gold standard identified 50 chest x-ray reports consistent with ALI and 46 that were not consistent with ALI in the 96-report set. The 857-report set included 392 chest x-ray reports that were consistent with ALI and 465 not consistent with ALI. The 11 reviewers’ agreement (96-report set) with the eventual gold standard is presented in Table 2.

Table 2. Reviewer agreement with the gold standard (96-report set)

Reviewer	**R**	**P**	**F**	**corr**
r1	0.94	0.98	0.96	0.917
r2	0.98	0.91	0.94	0.877
r3	0.80	0.95	0.87	0.762
r4	0.80	0.98	0.88	0.786
r5	0.62	1.00	0.77	0.601
r6	1.00	0.83	0.91	0.775
r7	0.92	0.94	0.93	0.771
r8	0.70	0.92	0.80	0.639
r9	0.70	1.00	0.82	0.671
r10	0.96	0.96	0.96	0.834
r11	0.92	0.98	0.95	0.813

R=Recall, P=Precision, F=F-measure, corr=Pearson correlation coefficient

We calculated recall, precision, F-measure and Pearson correlation coefficient for each reviewer's agreement with the final (majority vote based) gold standard. The F-measure for reviewers' gold standard agreement varied between 0.77 and 0.96. Precision and recall values had a range of 0.62 to 1.00. All the correlations with the gold standard are significant at 0.001 level.

3.3. List of keywords

The domain experts generated 48 phrases for the list of keywords. Table 3 shows some of the trigger phrases for the keyword search and their corresponding weights on the two scales. The weights represent the average opinion of two pulmonary care specialists on scale 1-3 and 1-10 developed independently.

Table 3. Example keywords and weights

Phrase	Weight - 3	Weight - 10
edema	2.5/3	8/10
lung edema	2.5/3	8/10
lung opacities	2/3	5.5/10
lung infiltrates	2/3	5.5/10
ALI	3/3	10/10
interstitial edema	1.5/3	4.5/10
pulmonary edema	3/3	9/10
both lungs	3/3	9/10

3.4. Keyword and weight based classification

We used the 48 phrases that domain experts generated and the corresponding weights to classify the chest x-ray reports in our first approach for the task. Table 4 shows the results of the three rounds of classification (raw keywords, 3-point and 10-point scale based). Results for the two corpora are presented separately.

Table 4. Keyword and weight based results

Corpus	R	P	F	corr
96-raw	0.88	0.83	0.85	0.687
96-w3	0.82	0.85	0.84	0.667
96-w10	0.72	0.88	0.80	0.617
857-raw	0.95	0.67	0.79	0.586
857-w3	0.91	0.71	0.80	0.599
857-w10	0.83	0.80	0.82	0.657

R=Recall, P=Precision, F=F-measure, corr==Pearson correlation coefficient

On both corpora the recall ranged between 0.72 and 0.95 while precision was between 0.67 and 0.88. The lowest value for F-measure was 0.79 and the highest was 0.85. All correlations of system outputs with gold standard were statistically significant at 0.001 level.

3.5. Machine learning based classification

In Table 5. we present the statistics for word unigram (-word) and 1-gram to 6-gram (-n1 to –n6) feature based systems for both the 96-report and 857-report corpora. In case of the character n-grams we used only the particular n-gram in the classification. For example, in case of 6-gram we used only the 6-grams and not other n-grams in the classification.

Table 5. MaxEnt based results

System	R	P	F	corr
96-word	0.83	0.78	0.80	0.621
96-n1	0.62	0.58	0.60	0.257*
96-n2	0.67	0.81	0.73	0.525
96-n3	0.82	0.85	0.84	0.637
96-n4	0.85	0.97	0.91	0.760
96-n5	0.77	0.73	0.75	0.554
96-n6	0.84	0.82	0.83	0.696
857-word	0.79	0.81	0.80	0.638
857-n1	0.73	0.71	0.72	0.366
857-n2	0.78	0.80	0.79	0.586
857-n3	0.81	0.79	0.80	0.628
857-n4	0.83	0.77	0.80	0.628
857-n5	0.86	0.85	0.86	0.669
857-n6	0.91	0.90	0.91	0.689

R=Recall, P=Precision, F=F-measure, corr=Pearson correlation coefficient, *=p-value >= 0.001

The word unigram based system performed equally for both corpora, with 0.80 F-measures. The character n-grams had a recall range of 0.62 to 0.85 and precision range of 0.58 to 0.97. The F-measures varied between 0.60 and 0.91 for the smaller corpus.

The 857-report set had recalls between 0.73 and 0.91, precision between 0.71 and 0.90 with F-measures between 0.72 and 0.91. The correlation with the gold standard for the character unigram had a p-value of 0.01 (marked with asterisk). All other correlations were significant at 0.001 level.

Table 6 shows the findings for the ROC statistics.

Table 6. ROC statistics for the 857-report set

System	Area	L95%	U95%
857-raw	0.782	0.751	0.813
857-w3	0.796	0.765	0.827
857-w10	0.829	0.800	0.859
857-word	0.818	0.788	0.849

Table 6. cont.

857-n1	0.683	0.646	0.719
857-n2	0.793	0.761	0.824
857-n3	0.815	0.784	0.845
857-n4	0.816	0.786	0.846
857-n5	0.835	0.806	0.864
857-n6	0.844	0.815	0.873

L95%=Lower bound, U95%=Upper bound of 95% Confidence Interval

Figure 1 shows that the peak for accuracy is at character 6-gram.

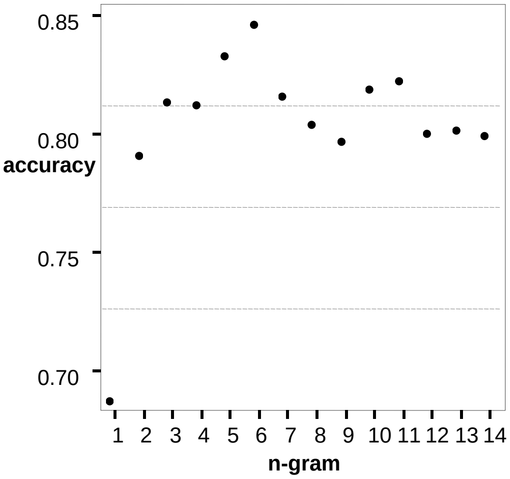

Figure 1. Accuracy by n-gram

4. Discussion

4.1. Interpretation of the results

Our study's major finding is that a Maximum Entropy and character 6-gram-based supervised chest x-ray classification system can achieve a higher F-value for ALI classification than keyword and simple heuristics-based systems. We found that the machine learning approach using character 6-grams achieved better F-measure than the raw keyword algorithm.

The 95% confidence intervals for the ROC AUC did not overlap for the character 6-gram and the raw keyword-based algorithm, indicating that the higher performance is not due to chance. While numerically higher, the ROC AUC 95% confidence intervals for the character 6-gram overlapped with the weight-based systems. Additional statistical methods to test the difference between keyword and machine learning-based approaches are needed. While there have been numerous studies in the general NLP field showing the advantages of machine learning over rule-based systems, to our best knowledge our project is the first that used machine learning for ALI classification. We demonstrated that the MaxEnt-based system not only produced good results but also did not require experts' input in generating a list of keywords.

It should be noted that, in contrast to the larger corpus where 6-gram-based approaches appear superior, in the 96-report corpus we found that a machine learning-based algorithm using character 4-grams achieved the numerically highest F-measure (with equal balance for recall and precision). This finding provides some additional support for the assertion that machine-learning-based approaches are superior to keyword-based approaches but highlights that further work is needed to clarify the ideal size of character n-gram to be used for classification of ALI.

Another potential advantage of the machine learning-based approach is that it is easier to set up than a keyword based pipeline. The stochastic classifier does not need domain expertise from the developers. The learner needs labeled training data but does not need experts to generate a keyword dictionary as the rule-based classifier does.

We have focused our discussion on results obtained from the larger corpus because we think it better represents the population of chest x-ray reports to which ALI classification algorithms would be most useful to apply. Although we obtained similar results for both corpora, as mentioned above, the character n-grams with the highest F-measures differed between the two corpora.

This difference could be due to the major difference in sample sizes and/or differences in enrollment approaches. We also observed that in the larger corpora there was more than one local optimum (Figure 1). This may be due to the fact we used n-grams only instead of using grams inclusive of all values up to "n". Again, future work will need to focus on identifying optimal character n-grams in the most relevant clinical populations.

4.2. Limitations and future research directions

Our study has several limitations. First, two corpora were prepared with slightly different selection and gold standard development criteria. The 96-report corpus was annotated by 11 annotators and the standard came from majority vote. The larger corpus was developed by only one annotator. We currently treat them as two separate corpora and build separate classifiers for each of them. In the future, we plan to have the larger corpus be annotated by more experts. We will also explore different methods of combining the data in the two corpora.

Second, we have not tested our algorithms on corpora used by other ALI research teams such as the ones discussed in the next section. This will be necessary to directly compare our systems. Such validation and direct comparisons are planned.

Third, in this study we use only character n-grams restricted to a specific "n". In future work we will use the n-grams inclusive of all grams up to "n". We also plan to find better features in general and better features for context. Our opinion is that the 6-gram achieved higher performance than the word unigram because the character 6-gram tapped into the potential of contextual features (we present the specific 6-grams in other venue). We think that we can develop context features beyond the potential of character grams.

5. Previous Work

We are aware of only two research groups that included chest x-ray report classification in their ALI surveillance system. Both research teams developed rule-based systems without a machine learning component and were focused on ALI screening rather than natural language processing research. Given this focus, their reports did not provide any detailed analysis or description of the text processing modules and, thus, it is difficult to draw direct comparisons between our approaches.

Herasevich et al. at the Mayo Clinic in Rochester, Minnesota aimed to determine the accuracy of computerized syndrome surveillance for detection of ALI and compared it with routine clinical assessment [7]. They included a free text Boolean query containing trigger words: ("bilateral" AND "infiltrate") OR "edema" in their ALI diagnosis screener. While this algorithm appears to have achieved good specificity, they did not provide any separate recall and precision statistics for the Boolean query module to allow a more direct comparison to our work.

Azzam et al. at the University of Pennsylvania designed an automated electronic system to screen for ALI in mechanically ventilated patients [8]. Similarly to Herasevich, their objective was the development of an ALI diagnostic tool and not an x-ray report classification system for ALI. However, their keyword search algorithm was more complex than Herasevich et al. Their report did not contain the keyword-based document classification statistics and so it is hard to extrapolate how their algorithm might have performed in our population.

6. Conclusion

We have shown that for the classification of ALI chest x-ray reports, using a machine learning approach with character 6-grams can achieve superior F-measures compared with a keyword-based system and achieves performance similar to the highest-performing physician annotators. This finding suggests that machine learning-based text classification modules might be used in the clinical setting to classify ALI with higher recall, precision and specificity than rule-based systems. Because the character n-gram based machine learning approach does not require hand crafted rules it is easier to scale to other disease categories than the keyword based systems.

7. References

[1] Erickson, S. E., Martin, G. S., Davis, J. L., Matthay, M. A., Eisner, M. D., "Recent trends in acute lung injury mortality: 1996-2005." *Critical Care Medicine,* 37: 1574-1579.

[2] Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., Buchanan, B. G. "A simple algorithm for identifying negated findings and diseases in discharge summaries." *Journal of Biomedical Informatics,* 2001, 34: 301-310.

[3] NegEx, Negation identification for clinical conditions, Retrieved August 01, 2009, from http://code.google.com/p/negex .

[4] Ratnaparkhi A. "A simple introduction to Maximum Entropy models for Natural language Processing." *IRCS Reports 97—08*, 1997, University of Pennsylvania.

[5] MALLET, Machine Learning for Language Toolkit, Retrieved August 02, 2009, from http://mallet.cs.umass.edu/.

[6] SPSS for Windows, Rel. 13.0. 2004. Chicago: SPSS Inc.

[7] Herasevich, V., Yilmaz, M., Khan H., Hubmayr R. D., Gajic, O. "Validation of an electronic surveillance system for acute lung injury." *Intensive Care Medicine,* 2009, 35; 1018-1023.

[8] Azzam, H. C., Khalsa, S. S., Urbani, R., Shah, C. V., Christie, J. D., Lanken, P. N., Fuchs, B. D. "Validation study of an automated electronic acute lung injury screening tool. " *Journal of the American Medical Informatics Association,* 2009, 16: 503-508.