Modeling Annotator Rationales with Application to Pneumonia Classification

Michael Tepper[1], Heather L. Evans[3], Fei Xia[1,2], Meliha Yetisgen-Yildiz[2,1]

[1]Department of Linguistics, [2]Biomedical and Health Informatics, [3]Department of Surgery
University of Washington
Seattle, WA 98195, USA
Email: {mtepper, hlevans, fxia, melihay}@uw.edu

Abstract

We present a technique to leverage annotator rationale annotations for ventilator assisted pneumonia (VAP) classification. Given an annotated training corpus of 1344 narrative chest X-ray reports, we report results for two supervised classification tasks: Critical Pulmonary Infection Score (CPIS) and the likelihood of Pneumonia (PNA). For both tasks, our training data contain annotator rationale snippets (i.e., spans of text that are relevant to annotator decisions). Because we assume that the snippet is not marked in the test data, we first built a sequential labeler to detect the location of snippets. The detected snippets are then used by the CPIS and PNA classifiers. Our experiments demonstrate that having access to detected annotator rationale leads to an incremental improvement in classification accuracy from 0.858 to 0.871 for CPIS, and from 0.785 to 0.821 for PNA.

1. Introduction

Free-text clinical reports, which are generated and stored during the course of patient care, often contain the richest, most detailed information about the patient's current health status. Ideally, we would like to enrich electronic medical records by leveraging the mass of information hidden in free text and connecting clinicians to critical (changes in) health status of every patient under their care.

The work we focus on in this paper involves finding actionable diagnostic information in narrative chest X-ray reports, part of a pneumonia phenotype detection system. Towards this goal, we have asked medical experts to annotate the chest X-ray reports with Clinical Pulmonary Infection Score (CPIS) to aid in making the pneumonia diagnosis (Xia and Yetisgen-Yildiz, 2012). We also asked the annotators to explicitly identify the rationale for their decisions by highlighting relevant portions of the text. The goal of this work is to show the effectiveness of the additional annotation in improving classification performance. Towards that end, we show that having access to annotator rationale leads to an incremental improvement in classification accuracy.

We also wanted to compare techniques for detecting annotator rationale. Such annotations are useful to practical users of clinical NLP technology (Yu et al., 2011). If an NLP technique could closely model the human annotator's selection of evidence, it would be a good indication of the feasibility of developing approaches that provide rationales as feedback. For this task, we were able to achieve performance with high recall and mixed accuracy.

2. Task and Annotation

Our recent work involves identifying ventilator-associated pneumonia (VAP) at an early stage. When a patient is supported with mechanical ventilation, it is essential that clinicians monitor for and respond to clinical signs of developing pneumonia. This is complicated by the fact that signs and symptoms of VAP are non-specific, defined criteria are subjective, and the presence of disease cannot be definitively established by a single test. Diagnosis of the disease involves inspecting and weighing multiple pieces of clinical information (i.e., interpreting chest radiographs, respiratory secretions and blood work), repeated at multiple points in time.

The task at hand involves analyzing one of the core test results necessary for making the diagnosis: the narrative chest X-ray report. We had annotators label each report for two factors that would be useful in making a diagnostic decision for pneumonia. The first factor is derived from the Clinical Pulmonary Infection Score (CPIS), which is widely used to assist clinicians in making a VAP diagnosis (Zilderberg, 2010). The radiographic feature of the CPIS has three associated labels: *1A "no infiltrate"*, *1B "diffuse infiltrate or atelectasis"*, and *1C "local infiltrate"*. The

second factor is the radiologist's suspicion of the patient having pneumonia (PNA), based only on information in that report. The possible labels for this factor include *2A "no suspicion of PNA"*, *2B "suspicion of PNA"*, and *2C "probable PNA"*.

After an initial round of annotation of 100 randomly selected documents, the annotators were asked to come together and, when they differed, defend their classification choices using rationale based on actual text in the reports. This discussion resulted a detailed annotation guideline for the task (Xia and Yetisgen-Yildiz, 2012). They then waited for a few days (so that they would be unlikely to remember the decisions on the 100 discussed reports) and re-annotated the reports based on the guidelines. Agreement level for the second round of annotation was κ=0.797 for CPIS and κ=0.697 for PNA. The rest of the documents were annotated by a single annotator. In total, 1344 reports were annotated by our medical experts.

In addition to CPIS and PNA labels, the annotators were also asked to mark rationale as text spans for their classification choices. For each text span, annotators could select from 3 possible category choices*: "evidence for CPIS"*, *"evidence for PNA"*, or *"evidence for both CPIS and PNA"*. Text spans annotated for *"PNA"* or *"CPIS and PNA"* support PNA classification decisions, while "*CPIS*" or *"CPIS and PNA"* support *CPIS* classification decisions. In Fig 1, the italicized text span, "*Patchy residual left upper lobe consolidation*", was annotated as *"CPIS and PNA"* so it supports the CPIS class label *1C localized infiltrate* and the PNA class label *2B suspicion of pneumonia*.

Out of 1344 reports, there were three examples where no CPIS label was indicated and one where no PNA label was indicated, leading to a training set of 1341 and 1343 documents for CPIS and PNA, respectively.

CHEST, PORTABLE 1 VIEW

INDICATION:

Intubated

COMPARISON: June 30, 2085

FINDINGS:

Lines and tubes are in an unchanged position

Cardiac and mediastinal contours are normal. Resolving pulmonary edema. *Patchy residual left upper lobe consolidation.* Small left pleural effusion. Bibasal atelectasis.

No new bony abnormality.

Figure 1. Sample Chest X-ray report with gold labels (CPIS:1C, PNA:2B). Rationale snippet shown in italicized text is evidence for both CPIS and PNA.

A summary of the dataset is provided in Table 1. The CPIS dataset was heavily biased towards *1B diffuse infiltrate or atelectasis*, and has very few rationale annotations for *1A no infiltrate* (negative class). The PNA dataset is more balanced, and has rationale annotations spread more evenly across the categories. Reports classified with negative categories *1A* and *2A* often had no descriptions of lung opacities or infiltrate, nor any other evidence for pneumonia; these reports received no rationale annotations from our annotators.

	Category	# Reports	# Reports w/rationale
CPIS	*1A: no infiltrate*	178 (0.13)	25 (0.02)
	1B: diffuse infiltrate or atelectasis	1011 (0.75)	1004 (0.75)
	1C: local infiltrate	152 (0.11)	152 (0.11)
		1341 (1.00)	1181 (0.88)
PNA	*2A: no suspicion of PNA*	856 (0.64)	362 (0.27)
	2B: suspicion of PNA	294 (0.22)	290 (0.22)
	2C: probable PNA	193 (0.14)	192 (0.14)
		1343 (1.00)	844 (0.63)

Table 1. Chest X-ray dataset statistics showing the number (percentage) of reports and number (percentage) of reports with rationale annotations by category.

3. Methodology

For our first foray into this task, we adopted a cascaded approach, first detecting rationale snippets, and then determining class labels based on the detected snippets. We chose this approach because in preliminary experiments we observed a marked improvement to overall classification accuracy, as well as F-score for some of the lower performing classes, when training and test data were limited strictly to the rationale snippets. For more discussion on these oracle results, see Section 5.2. Based on this observation, we hypothesized that similar improvement over baseline classification performance would be yielded so long as predicted rationale snippets retained enough of positive qualities that made the originals easier to classify. A flowchart of the system architecture is provided in Fig 2.

3.1 Rationale Snippet Prediction

Our annotators were instructed to highlight rationale snippets as any text string that was relevant to the classification decision. The resulting annotations agreed with sentence boundaries for the most part, at a rate of over 95%. The other 5% were highlighted at the sub-sentence level, typi-

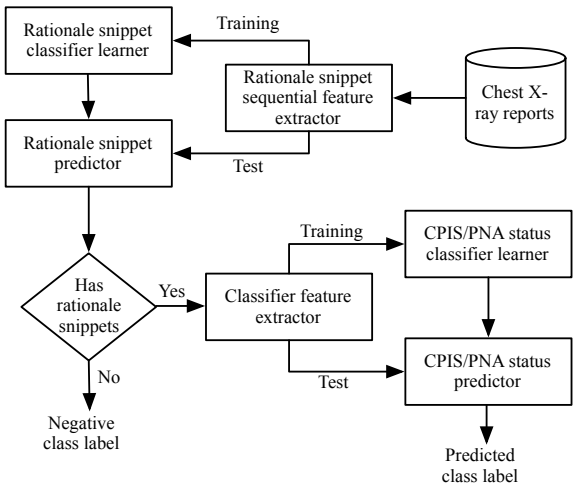

Figure 2. Flowchart of cascaded approach

cally along clausal boundaries. Because of the high agreement with sentence boundaries, we modeled rationale snippet prediction as a sentence-level sequential classification problem, though it could also be modeled as a more typical chunking task, at the word level. For both rationale types (evidence for CPIS and PNA), we compared standard *beginning, inside, outside* (BIO) to the reduced *inside, outside (*IO) tag set, in a maximum entropy sequential model.

We compared the benefit of using several token-level features (unigram, bigram) as well as features extracted from the tag sequence. When no tag sequence features were extracted, the approach is reduced to a binary classification for each sentence. Finally, we treated the labeling of CPIS and PNA rationale snippets as independent tasks, exploring the performance of each in parallel tagging experiments.

3.2 CPIS/Pneumonia Classification

For CPIS and Pneumonia classification, we trained a multiclass SVM[1] classifier using a bag-of-words approach. In our experiments, we compared extracting features from both rationale snippets and the whole document to extracting features from the rationale snippets only. For our baseline, we extracted features only from the whole document during training and test.

The feature templates we extracted are listed in Table 2. Since the purpose of this work was mainly to explore the benefits of rationale snippets on classification performance, there was very little focus on engineering complex features.

We ran two parallel sets of classification experiments treating the CPIS and PNA factors as independent classification tasks. When predicting based on rationale snippets, the CPIS classifier used snippets labeled as evidence for CPIS, while the PNA classifier used snippets labeled as evidence for PNA.

Feature	Description
Unigram	Unigrams (words occurring more than 3 times; stop words, punctuation, and digits filtered out)
Bigram	Bigrams (filtered the same way as for unigrams)
Concept	UMLS concept ID, as labeled by MetaMap (Aronson, 2001)
AlternateConj	Alternate proposing conjunction {*or, versus, VS*}

Table 2. Features used for classification

4. Previous work

Leveraging annotator rationales towards the goal of improving classification performance has been explored in recent work for the general domain (Zaidan et al., 2007; 2008) and the clinical domain (Yu et al., 2011).

Zaidan et al. (2007) developed a soft-margin SVM approach for sentiment analysis which was customized so that the objective function maximized an additional margin between actual training examples and a set of automatically generated pseudo-examples where a random subset of rationale strings are masked out. The advantage of this approach is it pushes predictive power onto maximally informative features from rationales, while still allowing other features to be learned from the remainder of the documents.

Yu et al. (2011) leveraged annotated rationale strings for concept identification primarily as a means of feature selection. For each concept, they took the top 30 most informative features comparing against the annotated rationale strings and some background text, and included those features within an existing concept identification system.

In these previous approaches, either the whole document was presumed potentially relevant to classification (Zaidan et al., 2007), or there was an existing system for generating a contextually relevant window/passage from input documents (Yu et al., 2011), which was not influenced by the rationale features. Our approach is different because we do not assume the whole document is relevant and we have no external means to determine relevance of a passage. We therefore use the snippets primarily as a way of determining passage relevance, rather than as a way of emphasizing informative features.

[1] In early experiments, we observed little performance difference between MaxEnt and SVM for this task; we chose SVM because of better training times.

		Sentence overlap			*Snippet overlap*		
		P	R	F1	P	R	F1
CPIS	BIO, w/prevTag feature	0.825	0.893	0.858	0.862	0.935	0.897
	IO, w/prevTag feature	**0.911**	0.840	0.874	**0.956**	0.899	0.927
	IO, no prevTag feature	0.866	**0.904**	**0.884**	0.916	**0.963**	**0.939**
PNA	BIO, w/prevTag feature	0.593	0.896	0.713	0.621	0.932	0.746
	IO, w/prevTag feature	0.612	**0.889**	**0.725**	**0.676**	**0.942**	**0.787**
	IO, no prevTag feature	**0.628**	0.802	0.704	0.668	0.940	0.781

Table 3. Results of rationale snippet prediction. For experiments w/prevTag we used beam search to find the best sequence.

5. Experiments

In this section, we present experiments on predicting rationale snippets (Section 5.1), followed by our experiments using rationale snippets to predict CPIS and pneumonia class labels (Section 5.2). As preprocessing for both 5.1 and 5.2, we used OpenNLP[2] to chunk reports into sentences and tokens.

5.1 Rationale Snippet Prediction

As mentioned in Section 3, we predicted rationale snippets for each factor type separately: (1) "evidence for CPIS", and (2) "evidence for PNA". In the gold annotations, we also have jointly labeled "evidence for both CPIS and PNA". For now, we have simply treated this label as marking independent snippets of both types: "evidence for CPIS" and "evidence for PNA", on the same text span. For evaluation of snippet prediction performance, we ran 5-fold cross validation on our training set of 1344 radiology reports for both factor types.

Evaluation measures

For evaluation, we calculated precision, recall, and F1 on two performance metrics. The first is a *sentence-overlap metric*, which counted a match whenever a sentence is marked as being (part of) a snippet by both the system and the gold standard. A sentence is counted as a false positive if it is marked as being (part of) a snippet by the system only, and counted as a false negative if it was marked as being (part of) a snippet by the gold-standard only. The second is a simple *snippet-overlap metric,* which counts a match whenever a snippet marked by the system overlaps with a gold snippet. It counts a false positive if a snippet marked by the system does not overlap with any gold standard snippet and a false negative if a gold-standard snippet does not overlap with any snippet marked by the system. The two measures are the same except that the former counts every sentence and the latter counts every snippet.

Experiment setup

The same bag-of-words feature template was used in all experiments, which included unigrams filtered to remove stop words, punctuation, and digits. Bigrams were tried, but they provided no benefit over unigrams alone, so these results are not presented. For all experiments we applied statistical feature selection and used the top 250 features, ranked in terms of information gain.

To improve recall, we trained only on the subset *S* of reports containing snippets. Training folds were selected only from *S*, while test folds were augmented by including the remaining (snippet-free) reports, distributed evenly across each fold. This selection technique always proved beneficial to overall performance, so we only report results with this technique turned on. We used the same 5 training and test folds for all experiments.

For modeling, we used the MALLET MaxEnt classifier (McCallum, 2002). For each experiment, we only used the top 250 features as ranked by information gain.

Results

Results for snippet prediction are given in Table 3. There was a large performance difference between CPIS and PNA snippet types. The best performing setting for CPIS measured 0.939 F1 on snippet-overlap, while the best setting for PNA measured just 0.787 F1 on snippet-overlap. This difference in F-score was caused by the relatively large difference in precision. Compared to CPIS snippets, the PNA rationale snippets tended to overgeneralize more from patterns observed in the training data. For example, many sentences that had references to lung opacities and infiltrates were falsely predicted as being (part of) a snippet, even if they provided little direct evidence for pneumonia. The resulting predicted snippets for PNA thus became more of a passage of potentially relevant text, than an actual rationale providing evidence for the annotation. The CPIS snippets also overgeneralized, but to a lesser extent.

Regarding the effect of tagging scheme, the BIO models did not perform as well as IO. This may be because, for

[2] OpenNLP. Available at: http://opennlp.apache.org/index.html

this task, there are few defining patterns to distinguish the first middle, or last sentence of a snippet. Therefore the main effect of splitting into B and I tags is simply to increase data sparsity.

Sequence features (prevTag) helped to improve precision for PNA snippet prediction, which led to an F-score improvement. There was also a precision improvement for CPIS snippets; however, the recall dropped, resulting in a negative effect on F-score. The best performing approach for CPIS was simple maximum entropy, where no sequence features were active. The biased dataset may have made sequence features less effective for CPIS snippets – with sequence features active, features for snippets supporting the minority class *1C local infiltrate* may have been more undertrained.

5.2 CPIS/Pneumonia Classification

For the following experiments, the rationale snippet prediction models were chosen as the best-performing models (one for CPIS, one for PNA) from the previous section. We used these models to predict the rationale, and then used the predicted rationale both in training and testing the classifier. For evaluation of classification performance, we ran 5-fold cross validation on our training set of 1341 radiology reports for CPIS, and 1343 radiology reports for PNA.

Experiment setup

We used the same bag-of-words feature template for all experiments, as given in Table 2. We do not present any results with bigrams or UMLS concepts, as they provided no benefit. Therefore, the only active features in all of our experiments were *Unigram* and *AlternateConj*.

For both CPIS and PNA tasks that used *only* the (oracle or predicted) rationale snippets, we restrict training data so that models are only trained on examples with snippets. At test time, any example that had no rationale snippets was assigned with the default negative category for the task (1A or 2A), as shown in the flowchart in Fig 2. Also, models that use predicted snippets are *trained* as well as *tested* on predicted snippets, as opposed to training on oracle snippets and testing on predicted snippets.

For modeling, we used the LIBSVM Java API (Chang and Lin, 2011), with the RBF kernel. We adjusted the soft-margin parameter (parameter C), selecting a higher value for the less noisy experiments that use only snippets features, and a lower value for the noisier experiments that use whole-document features.

Results

The results of our classification experiments are presented in Table 4. Classifiers that use predicted rationale snippets consistently outperform the baseline classifiers trained only on whole-document features. We observe an accuracy improvement from 0.858 to 0.871 for CPIS, and 0.785 to 0.821 for PNA, resulting in an error rate reduction of 9.1% and 16.7%, respectively.

For both the CPIS and PNA tasks, the rationale snippets features are most helpful when used on their own. When combined with whole-document features, performance decreases. This decrease is likely related to an increase in dimensionality of the feature space (adding many redundant features), and a procedural difference – in the experiments that only used snippets, documents without snippets were excluded from training and automatically classified as negative during testing.

To calculate the maximum potential benefit of the rationale snippets, we calculated oracle results where we trained and tested on the gold-standard rationale snippets from the training data. These results show that the CPIS factor has oracle accuracy of 0.919. Comparing that to the

		TP	FP	FN	P_{macro}	R_{macro}	$F1_{macro}$	Acc.
CPIS	Whole	1150	191	191	0.793	0.669	0.726	0.858
	Predicted Snippets	1168	173	173	0.789	**0.750**	**0.769**	**0.871**
	Whole + Predicted Snippets	1163	178	178	**0.798**	0.694	0.742	0.867
	Oracle Snippets	1232	109	109	0.880	0.839	0.859	0.919
	Whole + Oracle Snippets	1209	132	132	0.862	0.773	0.815	0.902
PNA	Whole	1054	289	289	0.702	0.666	0.684	0.785
	Predicted Snippets	1103	240	240	**0.754**	**0.726**	**0.740**	**0.821**
	Whole + Predicted Snippets	1080	263	263	0.751	0.672	0.709	0.804
	Oracle Snippets	1137	206	206	0.793	0.758	0.775	0.843
	Whole + Oracle Snippets	1118	225	225	0.771	0.754	0.762	0.832

Table 4. Results of CPIS and PNA classification. "Whole" indicates that features were extracted from the whole document, "Predicted Snippets" indicates that features were extracted from predicted rationale snippets, and "Oracle Snippets" indicates that features were extracted from the gold-standard rationale snippets. In "Whole+Snippets" experiments, snippet features were appended to whole-document features.

results with predicted snippets, this shows that CPIS/PNA classification results could be further improved if the performance of snippet prediction improves.

For CPIS, rationale-prediction model overgeneralization, which manifests as a tendency to predict all potentially relevant passages, rather than the specifically relevant passage for the instance in question, causes harm to classification performance. For example, there are many cases for CPIS class *2C: localized infiltrate* where the gold rationale snippet excludes a substring within the predicted snippet, e.g. "Increasing consolidation in the left lower lobe," but excluding the seemingly relevant "Unchanged atelectasis in the right middle lobe." In error analysis, we found that such passages in predicted snippets can push the classifier to learn less useful generalizations at training time, and make the wrong decisions at test time.

For PNA, even though the rationale-prediction model appears to overgeneralize in a worse manner, the overgeneralization appears to be less harmful to classification performance. We can potentially conclude from this that the main benefit of the PNA rationale snippets, both gold-standard and predicted, is to limit the context of classification decision to relevant information.

6. Discussion

In analyzing the errors made by the CPIS classifier, a frequent pattern emerges involving instances with strong evidence for **both** *1B: diffuse infiltrate or atelectasis*, and *1C: local infiltrate*. For many such instances, the correct answer will be *1C: local infiltrate*, but in the learned models, features for *1B: diffuse infiltrate or atelectasis* will be too powerful and tip the scales in that direction. In the future, we plan to apply a secondary technique, following Yu et al. (2011), to use gold-standard rationale to learn which features are most informative on an instance basis.

Another frequent issue for CPIS and PNA classification is the confusing signal introduced by hedges. Radiologists often hedge by listing multiple diagnoses as possibilities, separated by a conjunction like *'or'* or *'versus'*, after stating concrete observations. For example, a radiologist might write: "*Patchy consolidation right lung, likely pneumonia or atelectasis*." Here, the string "*likely pneumonia or atelectasis*" is a type of hedge. For future work, we would like determine if features built on hedge-detection have an impact on pneumonia-detection performance.

7. Conclusion

In this paper, we have shown an incremental improvement in classification performance when utilizing rationale annotation. We have devised a simple, sequential model for predicting the snippets, which has high recall, though it can have low precision. The effect of this snippet prediction model appears to be to identify relevant passages for the classification task, even if the passages may not be the most informative for a given instance.

There are other mechanisms for finding relevant passages of text, namely anchor or mention-based mechanisms, which use high-recall triggers to single out relevant passages in reports. A possible advantage to the current approach over a trigger-based approach is robustness to passages where triggers are not made explicit, because they are understood from the context. For example, in our chest X-ray reports, the concept *lung opacity* can be replaced with a vague descriptor like *haziness*, such as "haziness in right hemithorax." Under what conditions the current approach is preferable to a well-engineered set of triggers should be investigated in future work.

8. Acknowledgments

The work is partly supported by the University of Washington Research Royalty Fund, the Institute of Translational Health Sciences (UL1TR000423), UW K12 Comparative Effectiveness Research Training Program (K12 HS019482-01), and Agency for Healthcare Research and Quality. We would also like to thank anonymous reviewers for helpful comments.

References

Aronson, A.R. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proc. of AMIA Symposium*, 17–21.

Chang, C-.C., and Lin, C-.J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3):1–27.

McCallum, AK., 2002. A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu

Xia, F., and Yetisgen-Yildiz, M. 2012. Clinical corpus annotation: challenges and strategies. In *Proc. of the 3rd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with LREC 2012,*

Yu, S., Farooq, F., Krishnapuram, B., and Rao, B. Leveraging rich annotations to improve learning of medical concepts from clinical free text. In *Proc. of the ICML 2011 Workshop on Learning from Unstructured Clinical Text*

Zaidan, O., Eisner, J., and Piatko, C. 2007. Using "annotator rationales" to improve machine learning for text classification. In *Proc. of NAACL HLT*, 260–267.

Zaidan, O., and Eisner, J. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proc. of EMNLP, 31–40.*

Zilberberg M.D., and Shorr A.F. 2010. Ventilator-associated pneumonia: the clinical pulmonary infection score as a surrogate for diagnostics and outcome. *Clinical Infectious Diseases* 51(Suppl 1), S131-S135