Combining Structured and Free-text Data for Automatic Coding of Patient Outcomes

Suchi Saria[1], Gayle McElvain[2], Anand K. Rajani[3], Anna A. Penn[3], Daphne L. Koller[1]
[1]Computer Science Department, [2]Department of Linguistics, [3]Department of Pediatrics
Stanford University, Stanford CA 94305, USA
contact: ssaria@cs.stanford.edu

Abstract

Integrating easy-to-extract structured information such as medication and treatments into current natural language processing based systems can significantly boost coding performance; in this paper, we present a system that rigorously attempts to validate this intuitive idea. Based on recent i2b2 challenge winners, we derive a strong language model baseline that extracts patient outcomes from discharge summaries. Upon incorporating additional clinical cues into this language model, we see a significant boost in performance to F1 of 88.3 and a corresponding reduction in error of 23.52%.

Introduction

The modern hospital generates large volumes of data, which include discharge summaries, records of medicines administered, laboratory results and treatments provided. With the recent ubiquity of electronic medical record (EMR) databases, all of this patient information is often documented within a single storage system. Automated extraction of patient outcomes from this rich data source can serve as infrastructure for clinical trial recruitment, research, bio-surveillance and billing informatics modules. Previous works have harnessed state of the art natural language processing (NLP) techniques in extracting patient outcomes from discharge summaries [1-3]. Although these systems perform reasonably well, performance is limited by complex language structure in the dictated sentences. While majority of the current work is focusing on building increasingly sophisticated language models, we take a complementary approach to this problem by incorporating simple cues extracted from structured EMR data when available. For example, treatments and medications are prescribed by clinicians to specifically manage patient complications; thus, presence or absence of relevant treatments can provide independent indicators to disambiguate cases where current NLP approaches fail. Similarly, clinical events can also provide markers for specific complications.

Methods

Data Characteristics

We built and evaluated our system on the records of 275 premature infants born or transferred within the first week of life to the Stanford Lucile Packard Children Hospital's Neonatal Intensive Care Unit (NICU) after March 2008 and discharged before October 2009. We extracted discharge summaries, as well as laboratory reports of urine (188 reports) and blood cultures (590), radiology reports of ECHO (387) and head ultrasounds (534), medication events, and clinical events such as ventilator settings and tube placements. This study is approved under a Stanford IRB protocol.

Our task is to identify, for each infant, any complications that occurred during their length of stay in the hospital. Administrative data such as ICD9 codes are known to have poor granularity and accuracy for identifying patient outcomes [4,5]. To remedy this, two expert neonatologists formulated a list of all major complications observed in the NICU (Table 1). The data was annotated for these and any additional unlisted complications and subsequently reviewed by a team of three nurses and a physician. Overall, there were 628 unique complication-patient pairs marked as positive and 4872 complication-patient pairs marked as negative.

Language Features

In constructing a baseline language model over discharge summaries, our aim is to achieve the highest classification accuracy possible so as to accurately evaluate the incremental contribution of incorporating additional structured data from EMRs. Recent work has shown the success of rule-based models in this domain, in particular those employing hand-crafted string matching patterns to identify relevant lexical items and shallow semantic features [1,8]. While these models are not optimal on account of their inability to generalize, they usually have better performance than models which use general NLP strategies [11]. We modeled our language based feature set off the context-aware approach employed by the i2b2 Obesity Challenge winners, Solt et al [1]. This approach aims to identify and categorize typical linguistic contexts in which patient disease outcomes are mentioned. The types of contexts which suggest a positive, negative, or uncertain result are fairly consistent within the domain of medical records, making it possible to engineer regular expressions that capture and categorize a majority of these mentions correctly. Four basic types of language based feature comprise our baseline system.

Disease Mentions: In addition to complication /

disease names, this category includes patterns to capture abbreviations (e.g., *UTI* and *NEC*), alternate spellings (e.g., *haemorrhage* and *hemorrhage*), complication subclasses (e.g., *germinal matrix hemorrhage* and *intracranial hemorrhage* for IVH), and synonyms (e.g., *cardiac arrest* for arrhythmia.) Expert opinion was sought in increasing feature coverage, akin to querying UMLS. The deterministic model using just this set of rules maps most closely to the baseline binary classifier in [1].

Negations: We use a Negex inspired strategy to identify both sentential and noun-phrase negations that indicate a negative result pertaining to one of the above disease name mentions. General patterns such as *no|never MENTION* and *(no|without) evidence of MENTION* are used across all disease types, but disease specific negation patterns are also allowed where appropriate, e.g., *r/o SEPSIS*.

Uncertainty modifiers: Uncertain contexts are identified by patterns of similar construction to the negation patterns but include templates such as *(possible|suspected) MENTION* and *history of MENTION*. It is important for the system to identify regions of uncertainty in order to avoid overvaluing many disease name mentions. Disease specific uncertainty patterns may also be used to recognize information that is most likely unrelated to patient outcome, e.g., *family death* or *pregnancy related UTI*.

Correlated Words and phrases: This final category of language features came from reviewing with experts words that showed high correlation with the outcome label. Similar to the process of automatically extracting symptoms, medications, and related procedures from the description of ICD-9 codes, we reviewed our data with medical professionals and arrived at pattern matches for names and abbreviations of relevant antibiotics, treatments (*antibiotics discontinued* for sepsis ruled out), symptoms (*PAC* for arrhythmia) and tests (*head ultrasound*).

A total of 285 language features were extracted. We experimented with several ways of combining these language features in our baseline model; we delay this discussion to the results section.

Clinical features

Structured information in the patient EMR can be extracted from sources other than the discharge summary, including records from diagnostic tests, medication and treatments administered. We refer to such features as *clinical* features. These features were developed with guidance from a neonatologist in two half hour sessions. For each complication, we listed various treatment options, medications provided, diagnostic tests used or other clinical events that are synonymous with the complication. Table 1 lists the various classes of clinical features that were used for each complication. Our overarching principle in implementing clinical features was simplicity of extraction. While more fine-tuned models can be built to improve sensitivity/specificity of features extracted from these different sources, our experiments show that even these relatively simple features are enough to significantly improve performance of the overall system.

Medications (M): The EMR stores the medication name, dosage, along with the time at which the medication was administered as structured events. Rules of the form *(medication name(s), minimum length of prescription)* were obtained from the neonatologist for all relevant complications. Such a rule is activated if a medication in the rule is administered to the infant for at least the minimum time.

Clinical Events (E): For various clinical events associated with complications, we obtained rules of the form *(event name, minimum event duration, threshold event value)*. Events include therapies (for example, infants with respiratory distress syndrome are often on oxygen therapy) as well as lab measurement (for example, extended increase in creatinine measurements is indicative of a renal malfunction in infants).

Culture Reports (C): Culture status is relevant to various complications. A vast majority of the cultures have a section that summarizes the result of the culture, where "No growth" is mentioned unless any bacterial growth is observed. We note that the presence of growth may be a result of a contaminant, which is further discussed in the unstructured text section of the report. For our current study, we do not make this correction.

Radiology Reports (R): Our approach is based on prior work that placed second in a recent CMC challenge [8]. For each type of report, we extract sections in decreasing order of relevance until a non-empty section is available. The section is parsed for indications of the complication or symptom mentioned in a positive, negated or uncertain context using the language rules described earlier.

Learning Technique

For outcome label prediction, we use a penalized logistic regression model that combines all features. While a broad set of classifiers can be deployed, penalized logistic regression is known to perform well in the low data regime [6]. The weights for this model are learned using maximum likelihood regularized with ridge regression, which trades off fit to data with model complexity, as measured by the sum of the learned weights.

Complication	M	E	C	R
Respiratory Distress Syn (RDS)	X	X		
Sepsis	X			
Patent Ductus Arteriosus (PDA)	X			X
Bronchopulmonary Dyslapsia (BPD)	X	X		
Intraventricular Hemorrhage (IVH)				X
Died				
Pneumothorax (PNE)				
Adrenal Insufficiency (ADR)	X			
Coagnegative Stahylococcus (BCS)	X		X	
Necrotizing Enterocolitis (NEC)	X	X		
Bacterimia (BAC)	X		X	
Arrhythmia (ARR)	X			
Hydrocephalus (HYD)				X
Pulmonary Hemorrhage (PUL)				
Urinary Tract Infection (UTI)			X	
Adrenal Renal Failure (ARF)		X		
Pneumonia (PNA)				
Pulmonary Hypertension (PPHN)				
Seizure (SEI)	X			
Chronic Renal Failure (CRF)		X		

Table 1: List of complication-specific clinical features used. Complications are listed in order of decreasing frequency in our data set. Features are extracted from medications (M), clinical events (E), culture reports (C) and radiology reports (R). Overall, 33 clinical features are extracted.

That is, we optimize the training objective:

$$\arg\max_w \sum_{i=1:N} [-y_i w^T f_i + \ln(1+\exp(w^T f_i)) + \frac{1}{2\sigma^2}||w||^2$$

where N is the number of training examples; f_i and $y_i \in \{0,1\}$ are the features and label of the ith example, w is the weight vector, and σ controls the magnitude of the ridge penalty.

Similar to [7], we develop *transfer* features that represent patterns that repeat across multiple complications and allow us to generalize from one label to another without having seen mentions of that feature in the training data. For example, *without sepsis* and *without pneumonia* both suggest the mention of the disease in a negated context. With a transfer feature *without (disease name)*, a negative weight learned from sepsis is applied in the context of pneumonia. Other examples of transfer features include *(disease name) ruled out*, *concern for (disease name)*. Of particular interest is the feature *PosMention (infrequent disease name)* which encodes sharing only amongst infrequently occurring complications. Complications like sepsis that are rampant in the population are discussed in almost every discharge summary and are ruled out using tests. Infrequent complications are only discussed when the patients show complication-specific symptoms and thus, their mention alone is strongly correlated with having the complication. Each feature is encoded by a set of regular expressions that capture varying mentions in the data. Weight sharing was similarly introduced for clinical features that were common to multiple complications (e.g., a positive blood culture is a diagnostic test used for both BAC and BCS).

To learn the feature weights, in the training objective for each example we combine all the disease specific and transfer features that are activated. Thus, the inclusion of both transfer and disease specific features with a ridge penalty allows the model to learn specificity when there are large number of examples and generality for rare outcomes.

Results

We compute precision, recall, and F1 for each condition, and then compute overall precision, recall, and F1 using micro-averaging. All results reported are based on average test performance over 100 trials of randomized 70/30 train/test split. Significance values are computed using the bootstrap method on the 100 trials.

Baseline Language Model

Our aim in developing the language model (LM) was to maximize its performance, so as to best evaluate the incremental contribution obtained from the clinical features. Thus, the LM development was done on the entire dataset using random 70/30 train/test splits. The cross-validation parameter σ was set to 0.8 to optimize test performance of the LM in the hold-out set, and not subsequently adjusted for the inclusion of the clinical features.

We experimented with several approaches for combining the language features to derive a strong baseline (see Table 2). Similar to past winners [8], we experimented with pre-fixed weighting schemes. A hand-tuned model was derived as follows: for a given patient-complication pair, all sentences from the discharge summary that matched language features for that complication were extracted. Each sentence was allowed at most one vote; a "Yes" vote was assigned if only disease mentions without negations or uncertainty matched the sentence or a "No" vote if any negated mentions of the disease matched. To combine all votes, a model that counted "No" votes twice as much as "Yes" votes gave the best results. *DLM*, deterministic language model, shows the performance of this fixed weighting scheme model. *LLM*, learned language model, shows performance of the model with weights learned assuming the bag of all matched features using the learning technique described earlier. We also show contributions of component feature

classes to the baseline by adding them incrementally. We use the LLM (all features), with F1 of 84.7, as the baseline for comparison with the EMR model.

Model	Feature Set	Prec.	Recall	F1
DLM	All features	73.5	86.1	79.4
LLM	Disease Mentions	88.7	72.8	79.9
	+ Negations	90.7	78.2	83.9
	+ Uncertain	90.8	77.8	83.7
	+ Correlated	90.6	79.5	**84.7**

Table 2: Baseline: language model performance

Integrated EMR Model

The EMR model contains all language features as well as the clinical features. Unlike the language model, the clinical features did not have an iterative feature development phase and were determined apriori using expert medical knowledge. The model weights were trained using a bag of words assumption with weight sharing for the transfer features as detailed earlier. In Table 3, we report test performance of the EMR model against our best language model. Overall, the EMR model with average F1 score of 88.3 performs significantly (p-value = 0.007) better than the language model. Additionally, the complications for which the EMR model does not outperform are those for which there were no clinical features included. From Table 1, note that for each complication, clinical features were extracted from only one or two sources.

A post-hoc analysis of the results was done to understand the performance of our augmented model. We identify three distinct sources of error: (1) medical ambiguities, (2) feature error, i.e., failure of a language or clinical feature match on a specific instance, and (3) data extraction.

A significant source of error within the dataset is inherent ambiguity in the process of medical diagnosis. Beyond cases that are simply complex to code, there are patients for which even medical experts disagree about the underlying diagnosis. This is especially true in our patient population, who tend to have a multitude of secondary and tertiary complications stemming from their initial underlying condition. The highest achievable F1 score in our data with these examples included as errors is 96.3.

Feature errors in the language model (LM) can arise when context patterns fail to match because a lexical cue is separated from the disease mention by too much intervening text, but this turned out to be a relatively rare occurrence in our dataset. There were just four instances of error where syntactic parsing could have identified a modifier that was missed by regular expressions. A second type of language error, which

	Language Model			EHR Model		
Comp	Pr.	Re.	F1	Pr.	Re.	F1
RDS	96.2	93.8	95.0	96.8	94.5	**95.6**
SEPSIS	82.3	69.8	75.5	92.5	79.5	**85.5**
PDA	92.4	85.6	88.9	94.7	87.0	**90.7**
BPD	90.5	73.3	81.0	92.9	82.2	**87.2**
IVH	92.9	79.0	85.4	96.2	78.5	**86.5**
DIED	95.0	93.9	**94.5**	94.7	93.7	94.2
PNE	100.0	85.9	**92.4**	100.0	84.1	91.4
ADR	90.4	56.8	69.8	91.4	64.2	**75.4**
BCS	93.6	88.6	91.0	99.7	87.5	**93.2**
NEC	76.5	59.5	66.9	74.6	61.5	**67.4**
BAC	69.6	11.3	19.5	100.0	68.6	**81.3**
ARR	98.5	50.2	66.5	98.1	61.0	**75.2**
HYD	88.3	79.7	83.8	88.8	91.2	**90.0**
PUL	100.0	99.5	**99.8**	100.0	90.5	95.0
UTI	59.0	58.5	**58.7**	55.7	57.0	56.3
ARF	67.7	28.2	39.8	71.2	33.3	**45.4**
PNA	100.0	2.0	4.0	100.0	2.7	**5.3**
PPHN	58.3	59.6	58.9	58.6	60.3	**59.4**
SEI	54.8	43.8	48.6	60.9	48.6	**54.1**
ALL	90.6	79.5	84.7	93.5	83.6	**88.3**

Table 3: Performance comparison between the language model and the EMR model. For visual clarity, the winning model is bolded for each complication. Complications in gray are those for which no clinical features were available.

occurs mainly with our most frequent complications, SEPSIS and RDS, are spans that contain atypical contexts and/or require inference. In the sentence, *"The workup was entirely negative and antibiotics were discontinued in approximately four days"*, there is no explicit mention of the complication, yet we can infer the patient most likely underwent a workup for sepsis. The addition of our 'Correlated Words' rule set helps mitigate these errors. In this case, for example, the rule *antibiotics discontinued after X hrs/days* correctly matched. In the full model, there were five errors of this type for RDS, one for SEPSIS, and one for PDA. The final type of feature error in the LM model is the most common, with at least ten instances in our complete dataset. It results when multiple mentions of a disease occur in conflicting contexts throughout the document or even within a single sentence. Temporal event resolution might improve performance in such cases.

Feature errors can also arise in clinical features, although less frequently due to the simplicity of their extraction. Such errors do occur mainly because combinations not covered by our feature set were administered. For example, cefotaxime or vancomycin are administered for at least four days when a patient has sepsis. However, some patients were switched from one to the other midway through their course, a feature not

covered by our initial set.

A final source of error was due to errors in the data extraction software we used, which is still in the first cycle of development. For more than 10 patients, subsets of their clinical records such as ultrasound reports, culture reports or clinical events were missing in our extracted dataset. Furthermore, for textual reports, occasionally missing word boundaries resulted in feature match errors.

Overall, an improved clinical feature set with more coverage and better extraction software should bring performance much closer to the achievable F1-ceiling.

Discussion and Conclusion

In this paper, we present a system that rigorously validates an intuitive idea: integrating easy-to-extract structured information such as medications, treatments and laboratory results into current NLP-based information extraction systems can significantly boost coding accuracy. With the recent ubiquity of EMR systems, this data is broadly available in many contexts [9]. We believe this study opens several exciting avenues for future work.

Exploiting dependencies between the related tasks of predicting individual disease outcomes might improve performance; the application of Conditional Random Fields (CRFs)[10] towards this end would be an interesting extension to the current formulation. Richer features that encode dependencies between multiple features can also help improve precision. For example, the medication hydrocortisone can be given for many reasons; however, if it is administered soon after a cortisol stimulation test, then it is most likely given for adrenal insufficiency (ADR). Modeling such dependencies can improve feature specificities.

Our current implementation is limited by the need to obtain expert opinion similar to other rule-based systems. While rule-based systems have been very successful in recent challenges [11], they are more cumbersome to scale due to the information acquisition bottleneck. Moreover, there may be valuable rules that did not occur to the expert in the development cycle. To remedy this, in combination with existing medication indication dictionaries [12], techniques such as boosting [13] can be used to automatically construct candidate rules. Such feature induction can also be integrated into an interactive system that uses experts to evaluate proposed rules for medical plausibility.

Knowledge representation is a difficult and an open research area in NLP. Our system mitigates shortcomings of current NLP techniques by encoding additional independent sources of information that provide reinforcement where entirely language based systems err. This has the additional benefit of building a more comprehensive case for each patient providing the health experts with a transparent system where the evidence supporting each decision can be verified holistically.

Acknowledgements

We would like to thank N. Llerena for data extraction, P. Hartsell, J. Hall and B. Kogut for data annotation, D. Vickrey for insightful discussions, and R. Xu, M. Musen and the reviewers for feedback on the paper. S. Saria was supported by a Rambus Corporation Stanford Graduate fellowship. This work was supported with a CS department seed grant.

References

[1] Solt I, Tikk D, Gl V, Kardkovcs ZT. Semantic Classification of Diseases in Discharge Summaries Using a Context-aware Rule-based Classifier. J Am Med Inform Assoc. 2009.

[2] Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004.

[3] Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting Information from Textual Documents in the EHR: A Review of Recent Research. In IMIA Yearbook of Medical Informatics 2008.

[4] Campbell JR, Payne TH. A comparison of four schemes for codifications of problem lists. Proc Annu Symp Comput Appl Med Care. 1994.

[5] Solti I, Aaronson B, Fletcher G et al. Building an Automated Problem List Based on Natural Language Processing: Lessons Learned in the Early Phase of Development. AMIA Annu Symp Proc. 2008.

[6] Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. Biostatistics 2004.

[7] Crammer K, Dredze M, Ganchev K, Talukdar PP, Carroll S. Automatic Code Assignment to Medical Text. Proceedings of the Workshop on BioNLP 2007.

[8] Goldstein I, Arzumtsyan A, Uzuner O. Three Approaches to Automatic Assignment of ICD-9-CM Codes to Radiology Reports. AMIA Annu Symp Proc. 2007.

[9] Li L, Chase HS, Patel C et al. Comparing ICD9-Encoded Diagnoses and NLP-Processed Discharge Summaries for Clinical Trials Pre-Screening: A Case Study. AMIA Annu Symp Proc. 2008.

[10] Lafferty L, McCallum A, Pereira F. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. ICML, 2001.

[11] Uzuner O. Recognizing obesity and Co-morbidities in sparse data. J Am Med Inform Assoc 2009.

[12] Burton MM, Simonaiis L, Schadow G. Medication and Indication Linkage: A Practical Therapy for the Problem List? AMIA Annu Symp Proc. 2008.

[13] Friedman JH, Hastie T, Tibshirani R. Additive Logistic Regression: a Statistical View of Boosting. Technical Report, Dept. of Statistics, Stanford University, 1998.