Distinguishing Historical from Current Problems in Clinical Reports—Which Textual Features Help?

Danielle L. Mowery MS, Henk Harkema PhD, John N. Dowling MS MD,
Jonathan L. Lustgarten PhD, Wendy W. Chapman PhD
Department of Biomedical Informatics
University of Pittsburgh, Pittsburgh, Pa 15260, USA
dlm31@pitt.edu, heh23@pitt.edu, dowling@pitt.edu, jll47@pitt.edu, wec6@pitt.edu

Abstract

Determining whether a condition is historical or recent is important for accurate results in biomedicine. In this paper, we investigate four types of information found in clinical text that might be used to make this distinction. We conducted a descriptive, exploratory study using annotation on clinical reports to determine whether this temporal information is useful for classifying conditions as historical or recent. Our initial results suggest that few of these feature values can be used to predict temporal classification.

1 Introduction

Clinical applications for decision support, biosurveillance and quality of care assessment depend on patient data described in unstructured, free-text reports. For instance, patient data in emergency department reports contain valuable indicators for biosurveillance applications that may provide early signs and symptoms suggestive of an outbreak. Quality assurance departments can use free-text medical record data to assess adherence to quality care guidelines, such as determining whether an MI patient was given an aspirin within twenty-four hours of arrival. In either application, one must consider how to address the question of time, but each of the applications requires a different level of temporal granularity: the biosurveillance system needs a coarse-grained temporal model that discerns whether the signs and symptoms are historical or recent. In contrast, the quality assurance system needs a fine-grained temporal model to identify the admission event, when (or if) aspirin was given, and the order and duration of time between these events. One important problem in natural language processing is extracting the appropriate temporal granularity for a given task.

Many solutions exist for extracting temporal information, and each is designed to address questions of various degrees of temporal granularity, including determining whether a condition is historical or recent, identifying explicit temporal expressions, and identifying temporal relations among events in text. (Chapman et al., 2007; Zhou et al., 2008; Irvine et al., 2008; Verhagen and Pustejovsky, 2008; Bramsen et al., 2006). We previously extended the NegEx algorithm in ConText, a simple algorithm that relies on lexical cues to determine whether a condition is historical or recent (Chapman et al., 2007). However, ConText performs with moderate recall (76%) and precision (75%) across different report types implying that trigger terms and simple temporal expressions are not sufficient for the task of identifying historical conditions.

In order to extend work in identifying historical conditions, we conducted a detailed annotation study of potentially useful temporal classification features for conditions found in six genres of clinical text. Our three main objectives were: (1) characterize the temporal similarity and differences found in different genres of clinical text; (2) determine which features successfully predict whether a condition is historical, and (3) compare ConText to machine learning classifiers that account for this broader set of temporal features.

2 Temporality in Clinical Text

For several decades, researchers have been studying temporality in clinical records (Zhou and Hripcsak, 2007). Readers use a variety of clues to distinguish temporality from the clinical narrative, and we wanted to identify features from other tem-

Proceedings of the Workshop on BioNLP, pages 10–18,
Boulder, Colorado, June 2009.

poral models that may be useful for determining whether a condition is historical or recent.

There are a number of automated systems for extracting, representing, and reasoning time in a variety of text. One system that emerged from the AQUAINT workshops for temporal modeling of newspaper articles is TARSQI. TARSQI processes events annotated in text by anchoring and ordering them with respect to nearby temporal expressions (Verhagen and Pustejovsky, 2008). A few recent applications, such as TimeText and TN-TIES (Zhou et al., 2008; Irvine et al., 2008), identify medically relevant events from clinical texts and use temporal expressions to order the events. One method attempts to order temporal segments of clinical narratives (Bramsen et al., 2006). One key difference between these previous efforts and our work is that these systems identify all temporal expressions from the text and attempt to order all events. In contrast, our goal is to determine whether a clinical condition is historical or recent, so we focus only on temporal information related to the signs, symptoms, and diseases described in the text. Therefore, we ignore explicit temporal expressions that do not modify clinical conditions. If a condition does not have explicit temporal modifiers, we still attempt to determine the historical status for that condition (e.g., "Denies cough"). In order to improve the ability to determine whether a condition is historical, we carried out this annotation study to identify any useful temporal information related to the clinical conditions in six clinical genres. Building on work in this area, we explored temporal features used in other temporal annotation studies.

TimeML is a well-known standard for complex, temporal annotation. TimeML supports the annotation of events defined as "situations that happen or occur" and temporal expressions such as *dates* and *durations* in order to answer temporal questions about these events and other entities in news text (Saur´ı, et al., 2006). One notable feature of the TimeML schema is its ability to capture verb tense such as *past* or *present* and verb aspect such as *perfective* or *progressing*. We annotated verb tense and aspect in medical text according to the TimeML standard.

Within the medical domain, Zhou et al. (2006) developed an annotation schema used to identify temporal expressions and clinical events. They measured the prevalence of explicit temporal expressions and key medical events like *admission* or *transfer* found in discharge summaries. We used the Zhou categorization scheme to explore temporal expressions and clinical events across genres of reports.

A few NLP systems rely on lexical cues to address time. MediClass is a knowledge-based system that classifies the content of an encounter using both free-text and encoded information from electronic medical records (Hazelhurst et al., 2005). For example, MediClass classifies smoking cessation care delivery events by identifying the status of a smoker as *continued*, *former* or *history* using words like *continues*. ConText, an extension of the NegEx algorithm, temporally classifies conditions as historical, recent, or hypothetical using lexical cues such as *history*, *new*, and *if*, respectively (Chapman et al., 2007). Drawing from these applications, we used state and temporal trigger terms like *active*, *unchanged*, and *history* to capture coarse, temporal information about a condition.

Temporal information may also be implied in the document structure, particularly with regards to the section in which the condition appears. SecTag marks explicit and implicit sections found throughout patient H&P notes (Denny et al., 2008). We adopted some section headers from the SecTag terminology to annotate sections found in reports.

Our long-term goal is to build a robust temporal classifier for information found in clinical text where the output is classification of whether a condition is historical or recent (historical categorization). An important first step in classifying temporality in clinical text is to identify and characterize temporal features found in clinical reports. Specifically, we aim to determine which expressions or features are predictive of historical categorization of clinical conditions in dictated reports.

3 Historical Assignment and Temporal Features

We conducted a descriptive, exploratory study of temporal features found across six genres of clinical reports. We had three goals related to our task of determining whether a clinical condition was historical or recent. First, to develop a temporal classifier that is generalizable across report types, we compared temporality among different genres

of clinical text. Second, to determine which features predict whether a condition is historical or recent, we observed common rules generated by three different rule learners based on manually annotated temporal features we describe in the following section. Finally, we compared the performance of ConText and automated rule learners and assessed which features may improve the ConText algorithm.

Next, we describe the temporal features we assessed for identification of historical signs, symptoms, or diseases, including temporal expressions, lexical cues, verb tense and aspect, and sections.

(1) **Temporal Expressions:** Temporal expressions are time operators like dates (*May 5th 2005*) and durations (*for past two days*), as well as clinical processes related to the encounter (*discharge, transfer*). For each clinical condition, we annotated whether a temporal expression modified it and, if so, the category of temporal expression. We used six major categories from Zhou et al. (2006) including: *Date and Time*, *Relative Date and Time*, *Durations*, *Key Events*, *Fuzzy Time*, and *No Temporal Expression*. These categories also have types. For instance, *Relative Date and Time* has a type *Yesterday, Today* or *Tomorrow*. For the condition in the sentence "The patient had a stroke in ***May 2006***", the temporal expression category is *Date and Time* with type *Date*. Statements without a temporal expression were annotated *No Temporal Expression* with type *N/A*.

(2) **Tense and Aspect:** Tense and aspect define how a verb is situated and related to a particular time. We used TimeML Specification 1.2.1 for standardization of tense and aspect where examples of tense include *Past* or *Present* and aspect may be *Perfective*, *Progressive*, *Both* or *None* as found in Saur´ı, et al. (2006). We annotated the verb that scoped a condition and annotated its tense and aspect. The primary verb may be a predicate adjective integral to interpretation of the condition (Left ventricle ***is enlarged***), a verb preceding the condition (***has*** hypertension), or a verb following a condition (Chest pain ***has resolved***). In "her chest pain ***has resolved,"*** we would mark "has resolved" with tense *Present* and aspect *Perfective*. Statements without verbs (e.g., No murmurs) would be annotated *Null* for both.

(3) **Trigger Terms:** We annotated lexical cues that provide temporal information about a condition. For example, in the statement, "Patient has past ***history*** of diabetes," we would annotate "history" as *Trigger Term: Yes* and would note the exact trigger term.

(4) **Sections:** Sections are "clinically meaningful segments which act independently of the unique narrative" for a patient (Denny et al. 2008). Examples of report sections include *Review of Systems* (Emergency Department), *Findings* (Operative Gastrointestinal and Radiology) and *Discharge Diagnosis* (Emergency Department and Discharge Summary).

We extended Denny's section schema with explicit, report-specific section headers not included in the original terminology. Similar to Denny, we assigned implied sections in which there was an obvious change of topic and paragraph marker. For instance, if the sentence "the patient is allergic to penicillin" followed the *Social History* section, we annotated the section as *Allergies*, even if there was not a section heading for allergies.

4 Methods

4.1 Dataset Generation

We randomly selected seven reports from each of six genres of clinical reports dictated at the University of Pittsburgh Medical Center during 2007 These included Discharge Summaries, Surgical Pathology, Radiology, Echocardiograms, Operative Gastrointestinal, and Emergency Department reports. The dataset ultimately contained 42 clinical reports and 854 conditions. Figure 1 show our annotation process, which was completed in GATE, an open-source framework for building NLP systems (http://gate.ac.uk/). A physician board-certified in internal medicine and infectious diseases annotated all clinical conditions in the set and annotated each condition as either historical or recent. He used a general guideline for annotating a condition as historical if the condition began more than 14 days before the current encounter and as recent if it began or occurred within 14 days or during the current visit. However, the physician was not bound to this definition and ultimately used his own judgment to determine whether a condition was historical.

Provided with pre-annotated clinical conditions and blinded to the historical category, three of the authors annotated the features iteratively in groups of six (one of each report type) using guidelines we

developed for the first two types of temporal features (temporal expressions and trigger terms.) Between iterations, we resolved disagreements through discussion and updated our guidelines. Cohen's kappa for temporal expressions and trigger terms by the final iteration was at 0.66 and 0.69 respectively. Finally, one author annotated sections, verb tense, and aspect. Cases in which assigning the appropriate feature value was unclear were resolved after consultation with one other author-annotator.

4.2 Data Analysis

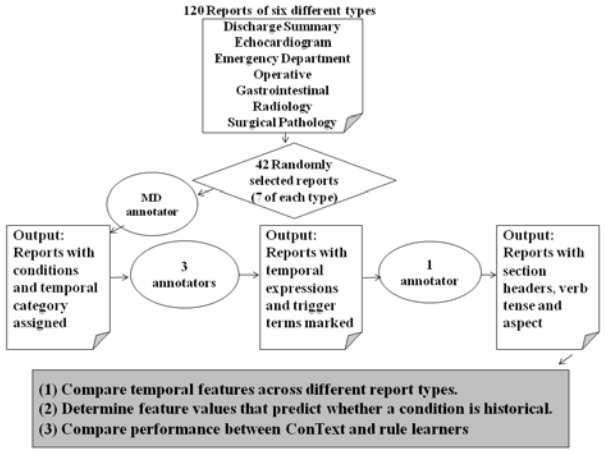

Figure 1. Annotation process for dataset and objectives for evaluation.

We represented each condition as a vector with temporal features and their manually-assigned values as input features for predicting the binary outcome value of historical or recent. We trained three rule learning algorithms to classify each condition as historical or recent: J48 Decision Tree, Ripper, and Rule Learner (RL) (Witten and Frank, 2005; Clearwater and Provost, 1990). Rule learners perform well at classification tasks and provide explicit rules that can be viewed, understood, and potentially implemented in existing rule-based applications. We used Weka 3.5.8, an openly-available machine learning application for prediction modeling, to implement the Decision Tree (J48) and Ripper (JRip) algorithms, and we applied an in house version of RL retrieved from www.dbmi.pitt.edu\probe. For all rule learners, we used the default settings and ran ten-fold cross-validation. The J48 algorithm produces mutually exclusive rules for predicting the outcome value. Thus, two rules cannot cover or apply to any one case. In contrast, both JRip and RL generate non-mutually-exclusive rules for predicting the outcome value. Although J48 and JRip are sensitive to bias in outcome values, RL accounts for skewed distribution of the data.

We also applied ConText to the test cases to classify them as historical or recent. ConText looks for trigger terms and a limited set of temporal expressions within a sentence. Clinical conditions within the scope of the trigger terms are assigned the value indicated by the trigger terms (e.g., historical for the term *history).* Scope extends from the trigger term to the end of the sentence or until the presence of a termination term, such as *presenting.* For instance, in the sentence "***History*** of CHF, ***presenting*** with chest pain," CHF would be annotated as historical.

5 Evaluation

To characterize the different reports types, we established the overall prevalence and proportion of conditions annotated as historical for each clinical report genre. We assessed the prevalence of each feature (temporal expressions, trigger terms, tense and aspect, and sections) by report genre to determine the level of similarity or difference between genres. To determine which features values are predictive of whether a condition is historical or recent, we observed common rules found by more than one rule learning algorithm. Amongst common rules, we identified new rules that could improve the ConText algorithm.

We also measured predictive performance with 95% confidence intervals of the rule learners and ConText by calculating overall accuracy, as well as recall and precision for historical classifications and recall and precision for recent classifications. Table 1 describes equations for the evaluation metrics.

Table 1. Description of evaluation metrics. RLP = rule learner prediction. RS = Reference Standard

	Historical		**Recent**	
	RLP	RS	RLP	RS
True Pos (TP)	Historical	Historical	Recent	Recent
False Pos (FP)	Historical	Recent	Recent	Historical
True Neg (TN)	Recent	Recent	Historical	Historical
False Neg (FN)	Recent	Historical	Historical	Recent

Recall: $\frac{\text{number of TP}}{\text{(number of TP + number of FN)}}$

Precision: $\frac{\text{number of TP}}{\text{(number of TP + number of FP)}}$

Accuracy: $\frac{\text{number of instances correctly classified}}{\text{total number of possible instances}}$

6 Results

Overall, we found 854 conditions of interest across all six report genre. Table 2 illustrates the prevalence of conditions across report genres. Emergency Department reports contained the highest concentration of conditions. Across report genres, 87% of conditions were recent (741 conditions). All conditions were recent in Echocardiograms, in contrast to Surgical Pathology reports in which 68% were recent.

Table 2. Prevalence and count of conditions by temporal category and report genre. DS = Discharge Summary, Echo = Echocardiogram, ED = Emergency Department, GI = Operative Gastrointestinal, RAD = Radiology and SP = Surgical Pathology. (%) = percent; Ct = count.

	Historical		Recent		
Report	(%)	Ct	(%)	Ct	Total Conditions
DS	(19)	38	(81)	158	196
Echo	(0)	0	(100)	199	199
ED	(17)	61	(83)	301	362
GI	(9)	3	(91)	32	35
RAD	(6)	2	(94)	32	34
SP	(32)	9	(68)	19	28
Total Conditions		113		741	854

6.1 Prevalence of Temporal Features

Table 3 shows that most conditions were not modified by a temporal expression or a trigger term. Conditions were modified by a temporal expression in Discharge Summaries more often than in other report genres. Similarly, Surgical Pathology had the highest prevalence of conditions modified by a trigger term. Operative Gastrointestinal and Radiology reports showed the lowest prevalence of both temporal expressions and trigger terms. Neither temporal expressions nor trigger terms occurred in Echocardiograms. Overall, the prevalence of conditions scoped by a verb varied across report types ranging from 46% (Surgical Pathology) to 81% (Echocardiogram).

Table 3. Prevalence of conditions modified by temporal features. All conditions were assigned a section and are thereby excluded. TE = temporal expression; TT = trigger term; V = scoped by verb.

	DS		Echo		ED		GI		RAD		SP	
	(%)	Ct	(%)	Ct	(%)	Ct	(%)	Ct	(%)	Ct	(%)	Ct
TE	(37)	73	(0)	0	(16)	59	(6)	2	(3)	1	(25)	7
TT	(25)	49	(0)	0	(19)	68	(9)	3	(3)	1	(39)	11
V	(70)	138	(81)	161	(62)	223	(69)	24	(76)	26	(46)	13

6.2 Common Rules

Rule learners generated a variety of rules. The J48 Decision Tree algorithm learned 27 rules, six for predicting conditions as historical and the remaining for classifying the condition as recent. The rules predominantly incorporated the trigger term and verb tense and aspect feature values. JRip learned nine rules, eight for classifying the historical temporal category and one 'otherwise' rule for the majority class. The JRip rules most heavily incorporated the section feature. The RL algorithm found 79 rules, 18 of which predict the historical category. Figure 2 illustrates historical rules learned by each rule learner. JRip and RL predicted the following sections alone can be used to predict a condition as historical: *Past Medical History*, *Allergies* and *Social History*. Both J48 and RL learned that trigger terms like *previous*, *known* and *history* predict historical. There was only one common, simple rule for the historical category found amongst all three learners: the trigger term *no change* predicts the historical category. All algorithms learned a number of rules that include two features values; however, none of the compound rules were common amongst all three algorithms.

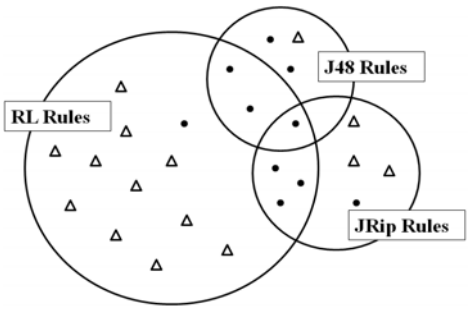

Figure 2. Historical rules learned by each rule learner algorithm. Black dots represent simple rules whereas triangles represent compound rules. Common rules shared by each algorithm occur in the overlapping areas of each circle.

6.3 Predictive Performance

Table 4 shows predictive performance for each rule learner and for ConText. The RL algorithm outperformed all other algorithms in almost all evaluation measures. The RL scores were computed based on classifying the 42 cases (eight historical) for which the algorithm did not make a prediction as recent. ConText and J48, which exclusively relied on trigger terms, had lower recall for the historical category.

All of the rule learners out-performed ConText. JRip and RL showed substantially higher recall for assigning the historical category, which is the most important measure in a comparison with ConText, because ConText assigns the default value of recent unless there is textual evidence to indicate a historical classification. Although the majority class baseline shows high accuracy due to high prevalence of the recent category, all other classifiers show even higher accuracy, achieving fairly high recall and precision for the historical cases while maintaining high performance on the recent category.

Table 4. Performance results with 95% confidence intervals for three rule learners trained on manually annotated features and ConText, which uses automatically generated features. Bolded values do not have overlapping confidence intervals with ConText. MCB = Majority Class Baseline (recent class)

Algorithm	Accuracy (Overall)	Recall (Historical)	Precision (Historical)	Recall (Recent)	Precision (Recent)
ConText	92.4	73.2	70.1	95.3	95.9
	90.8-94.4	70.5-76.4	67.2-73.4	94.1-96.1	94.8-97.5
J48	94.0	**62.8**	**88.8**	**98.8**	94.6
	92.6-95.8	**59.8-66.3**	**86.9-91.1**	**98.3-99.8**	93.3-96.3
JRip	**97.1**	**83.2**	**94.0**	**99.2**	97.5
	96.2-98.5	**80.9-85.9**	**92.6-95.8**	**98.8-100.0**	96.6-98.8
RL	**96.8**	**82.2**	**97.8**	**99.7**	97.5
	95.8-98.2	**79.9-85.0**	**97.0-99.0**	**99.5-100.0**	96.6-98.8
MCB	86.9	--	--	100.0	0.0

7 Discussion

Our study provides a descriptive investigation of temporal features found in clinical text. Our first objective was to characterize the temporal similarities and differences amongst report types. We found that the majority of conditions in all report genres were recent conditions, indicating that a majority class classifier would produce an accuracy of about 87% over our data set. According to the distributions of temporal category by report genre (Table 2), Echocardiograms exclusively describe recent conditions. Operative Gastrointestinal and Radiology reports contain similar proportions of historical conditions (9% and 6%). Echocardiograms appear to be most similar to Radiology reports and Operative Gastrointestinal reports, which may be supported by the fact that these reports are used to document findings from tests conducted during the current visit. Emergency Department reports and Discharge Summaries contain similar proportions of historical conditions (17% and 19% respectively), which might be explained by the fact that both reports describe a patient's temporal progression throughout the stay in the Emergency Department or the hospital.

Surgical Pathology reports may be the most temporally distinct report in our study, showing the highest proportion of historical conditions. This may seem counter-intuitive given that Surgical Pathology reports also facilitate the reporting of findings described from a recent physical specimen. However, we had a small sample size (28 conditions in seven reports), and most of the historical conditions were described in a single addendum report. Removing this report decreased the prevalence of historical conditions to 23% (3/13).

Discharge Summaries and Emergency Department reports displayed more variety in the observed types of temporal expressions (9 to 14 subtypes) and trigger terms (10 to 12 terms) than other report genres. This is not surprising considering the range of events described in these reports. Other reports tend to have between zero and three subtypes of temporal expressions and zero and seven different trigger terms. In all report types, temporal expressions were mainly subtype *past*, and the most frequent trigger term was *history*.

Our second objective was to identify which features predict whether a condition is historical or recent. Due to high prevalence of the recent category, we were especially interested in discovering temporal features that predict whether a condition is historical. With one exception (*date* greater than four weeks prior to the current visit), temporal expression features always occurred in compound rules in which the temporal expression value had to co-occur with another feature value. For instance, any temporal expression in the category *key event* had to also occur in the *secondary diagnosis* section to classify the condition as historical. For ex-

ample, in "SECONDARY DIAGNOSIS: Status post ***Coronary artery bypass graft*** with complication of mediastinitis" the key event is the *coronary artery bypass graft*, the section is *secondary diagnosis*, and the correct classification is historical.

Similarly, verb tense and aspect were only useful in conjunction with other feature values. One rule predicted a condition as historical if the condition was modified by the trigger term *history* and fell within the scope of a *present tense verb with no aspect*. An example of this is "The patient *is* a 50 year old male with ***history*** of hypertension." Intuitively, one would think that a past tense verb would always predict historical; however, we found the presence of a past tense verb with no aspect was a feature only when the condition was in the *Patient History* section. Sometimes the absence of a verb in conjunction with another feature value predicted a condition as historical. For example, in the sentences "PAST MEDICAL HISTORY: ***History*** of COPD. ***Also*** *diabetes…"* *also* functioned as a trigger term that extended the scope of a previous trigger term, *history*, in the antecedent sentence.

A few historical trigger terms were discovered as simple rules by the rule learners: *no change*, *previous*, *known*, *status post*, and *history*. A few rules incorporated both a trigger term and a particular section header value. One rule predicted historical if the trigger term was *status post* and the condition occurred in the *History of Present Illness* section. This rule would classify the condition *CABG* as historical in "HISTORY OF PRESENT ILLNESS: The patient is...***status post*** CABG." One important detail to note is that a number of the temporal expressions categorized as *Fuzzy Time* also act as trigger terms, such as *history* and *status post*—both of which were learned by J48. A historical trigger term did not always predict the category historical. In the sentence "No focal sensory or motor deficits on ***history***," *history* may suggest that the condition was not previously documented, but was interpreted as not presently identified during the current physical exam.

Finally, sections appeared in the majority of JRip and RL historical rules: 4/8 simple rules and 13/18 compound rules. A few sections were consistently classified as historical: *Past Medical History*, *Allergies*, and *Social History*. One important point to address is that these sections were manually annotated.

Our results revealed a few unexpected observations. We found at least two trigger terms indicated in the J48 rules, *also* and *status post*, which did not have the same predictive ability across report genres. For instance, in the statement "TRANSFER DIAGNOSIS: ***status post*** coiling for left posterior internal carotid artery aneurysm," *status post* indicates the reason for the transfer as an inpatient from the Emergency Department and the condition is recent. In contrast, *status post* in a Surgical Pathology report was interpreted to mean historical (e.g., PATIENT HISTORY: ***Status post*** double lung transplant for COPD.) In these instances, document knowledge of the meaning of the section may be useful to resolve these cases.

One other unexpected finding was that the trigger term *chronic* was predictive of recent rather than historical. This may seem counterintuitive; however, in the statement "We are treating this as *chronic* musculoskeletal pain with oxycodone", the condition is being referenced in the context of the reason for the current visit. Contextual information surrounding the condition, in this case treating or administering medication for the condition, may help discriminate several of these cases.

Our third objective was to assess ConText in relation to the rules learned from manually annotated temporal features. J48 and ConText emphasized the use of trigger terms as predictors of whether a condition was historical or recent and performed with roughly the same overall accuracy. JRip and RL learned rules that incorporated other feature values including sections and temporal expressions, resulting in a 12% increase in historical recall over ConText and a 31% increase in historical recall over J48.

Many of the rules we learned can be easily extracted and incorporated into ConText (e.g., trigger terms *previous* and *no change*). The ConText algorithm largely relies on the use of trigger terms like *history* and one section header, *Past Medical History*. By incorporating additional section headers that may strongly predict historical, ConText could potentially predict a condition as historical when a trigger term is absent and the header title is the only predictor as in the case of "ALLERGIES: peanut allergy". Although these sections header may only be applied to Emergency Department and Discharge Summaries, trigger terms and temporal expressions may be generalizable across genre of reports. Some rules do not lend themselves

to ConText's trigger-term-based approach, particularly those that require sophisticated representation and reasoning. For example, ConText only reasons some simple durations like *several day history*. ConText cannot compute dates from the current visit to reason that a condition occurred in the past (e.g., stroke in *March 2000).* The algorithm performance would gain from such a function; however, such a task would greatly add to its complexity.

8 Limitations

The small sample size of reports and few conditions found in three report genres (Operative Gastrointestinal, Radiology, and Surgical Pathology) is a limitation in this study. Also, annotation of conditions, temporal category, sections, verb tense and aspect were conducted by a single author, which may have introduced bias to the study. Most studies on temporality in text focus on the temporal features themselves. For instance, the prevalence of temporal expressions reported by Zhou et al. (2006) include all temporal expressions found throughout a discharge summary, whereas we annotated only those expressions that modified the condition. This difference makes comparing our results to other published literature challenging.

9 Future Work

Although our results are preliminary, we believe our study has provided a few new insights that may help improve the state of the art for historical categorization of a condition. The next step to building on this work includes automatically extracting the predictive features identified by the rule learners. Some features may be easier to extract than others. Since sections appear to be strong indicators for historical categorization we may start by implementing the SecTag tagger. Often a section header does not exist between text describing the past medical history and a description of the current problem, so relying merely on the section heading is not sufficient. The SecTag tagger identifies both implicit and explicit sections and may prove useful for this task. To our knowledge, SecTag was only tested on Emergency Department reports, so adapting it to other report genres will be necessary. Both JRip and RL produced high performance, suggesting a broader set of features may improve historical classification; however, because these features do not result in perfect performance, there are surely other features necessary for improving historical classification. For instance, humans use medical knowledge about conditions that are inherently chronic or usually experienced over the course of a patient's life (i.e., HIV, social habits like smoking, allergies etc). Moreover, physicians are able to integrate knowledge about chronic conditions with understanding of the patient's reason for visit to determine whether a chronic condition is also a recent problem. An application that imitated experts would need to integrate this type of information. We also need to explore adding features captured at the discourse level, such as nominal and temporal coreference. We have begun work in these areas and are optimistic that they will improve historical categorization.

10 Conclusion

Although most conditions in six clinical report genres are recent problems, identifying those that are historical is important in understanding a patient's clinical state. A simple algorithm that relies on lexical cues and simple temporal expressions can classify the majority of historical conditions, but our results indicate that the ability to reason with temporal expressions, to recognize tense and aspect, and to place conditions in the context of their report sections will improve historical classification. We will continue to explore other features to predict historical categorization.

Acknowledgments

This work was funded by NLM grant 1 R01LM009427-01, "NLP Foundational Studies and Ontologies for Syndromic Surveillance from ED Reports".

References

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. *Finding Temporal Order in Discharge Summaries.* AMIA Annu Symp Proc. 2006; 81–85

Wendy W Chapman, David Chu, and John N. Dowling. 2007. *ConText: An Algorithm for Identifying Contextual Features from Clinical Text.* Association for Computational Linguistics, Prague, Czech Republic

Scott H. Clearwater and Foster J. Provost. 1990. *RL4: A Tool for Knowledge-Based Induction.* Tools for Artificial Intelligence, 1990. Proc of the 2nd Intern IEEE Conf: 24-30.

Joshua C. Denny, Randolph A. Miller, Kevin B. Johnson, and Anderson Spickard III. 2008. *Development and Evaluation of a Clinical Note Section Header Terminology. SNOMED.* AMIA 2008 Symp. Proceedings: 156-160.

Brian Hazlehurst, H. Robert Frost, Dean F. Sittig, and Victor J. Stevens. 2005. *MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record.* J Am Med Inform Assoc 12(5): 517-29

Ann K. Irvine, Stephanie W. Haas, and Tessa Sullivan. 2008. *TN-TIES: A System for Extracting Temporal Information from Emergency Department Triage Notes.* AMIA 2008 Symp Proc: 328-332.

Roser Saur´ı, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. *TimeML Annotation Guidelines Version 1.2.1.* at: http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf

Marc Verhagen and James Pustejovsky. 2008. *Temporal Processing with TARSQI Toolkit.* Coling 2008: Companion volume – Posters and Demonstrations, Manchester, 189–192

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Li Zhou, Genevieve B. Melton, Simon Parsons and George Hripcsak. 2006. *A temporal constraint structure for extracting temporal information from clinical narrative.* J Biomed Inform 39(4): 424-439.

Li Zhou and George Hripcsak. 2007. *Temporal reasoning with medical data--a review with emphasis on medical natural language processing.* J Biomed Inform Apr; 40(2):183-202.

Li Zhou, Simon Parson, and George Hripcsak. 2008. *The Evaluation of a Temporal Reasoning System in Processing Discharge Summaries*. J Am Med Inform Assoc 15(1): 99–106.