Journal of Biomedical Informatics 45 (2012) 71–81

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Building an automated SOAP classifier for emergency department reports

Danielle Mowery [a,*], Janyce Wiebe [b], Shyam Visweswaran [a,b], Henk Harkema [a], Wendy W. Chapman [a,c]

[a] Department of Biomedical Informatics, University of Pittsburgh, Parkvale Building M-183, 200 Meyran Avenue, Pittsburgh, PA 15260, USA
[b] Intelligent Systems Program, University of Pittsburgh, 5113 Sennott Square, 210 South Bouquet Street, Pittsburgh, PA 15260, USA
[c] Division of Biomedical Informatics, University of California, San Diego, 9500 Gilman Dr., Bldg 2 #0728, La Jolla, CA 92093-0728, USA

ARTICLE INFO

Article history:
Received 20 June 2011
Accepted 31 August 2011
Available online 9 September 2011

Keywords:
Support vector machine
SOAP notes
Problem oriented medical record
POMR
Prediction

ABSTRACT

Information extraction applications that extract structured event and entity information from unstructured text can leverage knowledge of clinical report structure to improve performance. The Subjective, Objective, Assessment, Plan (SOAP) framework, used to structure progress notes to facilitate problem-specific, clinical decision making by physicians, is one example of a well-known, canonical structure in the medical domain. Although its applicability to structuring data is understood, its contribution to information extraction tasks has not yet been determined. The first step to evaluating the SOAP framework's usefulness for clinical information extraction is to apply the model to clinical narratives and develop an automated SOAP classifier that classifies sentences from clinical reports. In this quantitative study, we applied the SOAP framework to sentences from emergency department reports, and trained and evaluated SOAP classifiers built with various linguistic features. We found the SOAP framework can be applied manually to emergency department reports with high agreement (Cohen's kappa coefficients over 0.70). Using a variety of features, we found classifiers for each SOAP class can be created with moderate to outstanding performance with F_1 scores of 93.9 (*subjective*), 94.5 (*objective*), 75.7 (*assessment*), and 77.0 (*plan*). We look forward to expanding the framework and applying the SOAP classification to clinical information extraction tasks.

Published by Elsevier Inc.

1. Introduction

Health informatics applications aim to support workflow in clinical environments, provide decision and alert functions, detect disease outbreaks, ensure standard quality care practices, and facilitate biomedical research. One aspect of informatics research that has actively fostered enhancements to these application areas is natural language processing (NLP). NLP entails the use of computers to identify, encode and extract information from naturally written or spoken language in a structured form rendering it useful for use by other computerized applications. In recent years, NLP has been used to enhance disease outbreak detection [1], structure free-text as features for decision support [2], assist tracking of quality measures and guideline adherence [3], answer clinical questions, [4] and provide access to knowledge trapped in free-text resources such as clinical reports and biomedical literature [5]. Understanding the content of clinical narratives requires organizing and synthesizing clinical entities and events like conditions, procedures, and medications with contextual information indicating when such events occurred, in what order and who reported the information relative to the expected course of events within a patient visit. Successful automated natural language understanding requires integration of this broad spectrum of information. In this paper, we describe the development of an automated sentence classifier for the SOAP model—a framework used by clinicians to organize and interpret clinical information within a note—with the belief that this model can contribute to a larger model for understanding the course of a patient's hospital visit as described in a clinical report.

For half a century, the Subjective, Objective, Assessment, and Plan (SOAP) framework, a high-level model used to structure progress notes, has been a critical component in organizing clinical information in a way that supports assessment, reasoning, and decision making [6]. If the SOAP structure aids human interpretation, it may also be a useful component in an NLP application seeking to automatically structure and interpret information described in clinical reports. For instance, in studies on guideline adherence, it is critical to differentiate between medications a patient is already taking and medications that were newly prescribed at the visit or prescribed at discharge. Identification of the medication name is not sufficient, because the same medication name may occur in all three scenarios. Understanding the context within which the medication is described in the report is vital to knowing whether a medication is being taken currently ("MEDICATIONS: **Prilosec 20 mg** 1 tablet a day") or was prescribed at the visit

[*] Corresponding author. Fax: +1 412 647 7190.
E-mail address: dlm31@pitt.edu (D. Mowery).

1532-0464/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jbi.2011.08.020

(“Prescribed **Prilosec 20 mg**”). SOAP sentence classifications for these examples can help disambiguate existing (*subjective*) medications administered during the patient visit from planned (*plan*) medications prescribed for discharge.

Recently, researchers such as Solti et al., Cao et al., and Meystre and Haug have focused their efforts toward automatically condensing a diverse set of reports (*radiology, consult notes, H&P and discharge summaries, etc.*) into a concise problem list or clinical summary [7–10]. A SOAP classifier may increase the sensitivity of systems that identify and organize concepts into a problem list from ambulatory care reports to support staff communication during transition of patient care as advocated by the Center for Medicare and Medicaid Services (CMS), Joint Commission (JCAHO), and Institute of Medicine (IOM) [11–14]. For instance, Van Vleck et al. examines the structure necessary for generating a complete, problem list of preexisting conditions, procedures, and medication concepts from the past medical history sections of initial visit notes [15]. Preexisting concepts that are not specified in the past medical history section could be identified using sentences labeled as SOAP classes and the Unified Medical Language System (UMLS) concept tagging. Additionally, we would like to investigate how the SOAP framework can be used to organize the symptoms (S: *subjective*), signs (O: *objective*), reasonings (A: *assessment*) and treatments (P: *plans*) mentioned in the report relative to coded diagnoses (numbered conditions with this supporting clinical information) into a problem-oriented SOAP note from the free-text emergency department report.

2. Background

2.1. Document segmentation and sentence classification

Document segmentation and sentence classification are common techniques used in text processing. These techniques are applied in both non-medical and medical NLP domains such as newspaper articles, legal judgments, MEDLINE papers, and clinical texts.

Document segmentation is the application of statistical and rule-based methods to segment text into coherent, informative text blocks to filter and reduce the computational effort in information extraction and retrieval systems. Utiyama and Isahara tested a domain-independent document segmenter varying the number of segments using a statistical approach on a subset of the Brown Corpus [16]. Apostolova et al. train a document segmenter that classifies text segments into eight semantic units from radiology reports [17]. Denny et al. developed section taggers for labeling both explicit and implicit section changes using naïve Bayes scoring in history and physical exams [18]. Cho et al. use rule-based filters based on string, phrase, lexical and statistical analysis to tag sections in radiology and urology reports [19].

Sentence classification uses similar approaches to generate coherent text blocks or place concepts from a document in context or extract the main point from the text in document summaries. Bhowmick et al. trained Decision Tree learners (ADTboost.MH) to assign multiple emotion tags (disgust, fear, *etc.*) to sentences from news articles to improve human centered computation efforts using machines [20]. Hachey and Grover trained a variety of machine learning algorithms (naïve Bayes, Decision Tree, *etc.*) to classify sentences from legal judgments into a rhetorical scheme of 7 classes (fact, proceedings, background, *etc.*) to generate a “flexible summary” of the judgment [21]. Ruch et al. trained Bayesian, argumentative classifiers to tag sentences from MEDLINE abstracts into four-class categories (purpose, methods, results and conclusion) in an attempt to select a unique summary sentence from a study [22]. Kim et al. trained Conditional Random Fields to assign sentences to more than one “PICO” inspired classes (e.g., background, intervention, *etc.*) from MEDLINE abstracts that focused on randomized control trials [23].

2.2. Canonical sections can improve system performance

For some tasks, classifiers trained on canonical sections have improved system performance. Doan et al. derived a four-class annotation scheme (headline, lead, content and comment) for modeling section structure in news articles used to monitor the internet for indication of biological threats and outbreaks [1]. They found that the lead section resulted in the highest accuracy for identifying relevant texts for their system. Wang et al. found that incorporating knowledge of sections from clinical notes improved recall and precision scores for identifying symptoms manifested by diseases and for identifying adverse drug events [5]. The SOAP structure has the potential to improve performance in extracting existing medications and treatment plans in a system that aims to summarize medication administration throughout a patient visit or determine whether clinical protocols were followed. Additionally, the SOAP structure may increase the sensitivity of problem list generation by tagging sentences containing relevant concepts outside conventionally used sections (Past Medical History, Impression, Plan, *etc.*) from various report types (*discharge summaries, visit notes*, *etc.*) [7,12].

2.3. The SOAP model in medicine

The Subjective, Objective, Assessment and Plan (SOAP) framework was first developed by Larry Weed in 1969 to promote structuring of progress notes to support clinical reasoning and diagnosis in the problem-oriented medical record (POMR) [6]. In the POMR, clinicians number each symptom experienced by the patient. For each sign and symptom, the clinician lists four kinds of information (S) *subjective*, (O) *objective*, (A) *assessment* and (P) *plan*. The clinician begins by documenting symptoms to understand the patient’s clinical state (S). Next, the clinician records signs, quantifiable data and scientific evidence experienced by the patient (O). The clinician unifies and critically evaluates these prior pieces of information to formulate a working diagnosis (A) and, finally, reports the care plan to treat the underlying condition (P). We are currently developing NLP applications to model the clinical care practice in an emergency care setting and would like to incorporate the SOAP framework into a model for more accurate interpretation and understanding of events in the patient’s hospital visit.

2.3.1. Extensions of the SOAP framework

The SOAP framework and its derivatives remain a standard organizational tool for documentation generated in nursing, dentistry, psychiatry, occupational therapy and emergency medicine, among other domains [24–27]. Some definitions ground the interpretation of *subjective* and *objective* data elements as a matter of source where subjective information is elicited from the patient and objective information is provided by the medical professional. This is advocated in cases that the patient may not be a reliable source of information (e.g. a confused or drug-seeking patient). Other definitions place an equal emphasis on each of these pieces of information regardless of source and define *subjective* elements as a matter of time – historical or background information. The acronym for representing this structure could be History–Objective–Assessment–Plan (HOAP). Several suggestions have been made for the order the information is presented such as History–Objective–Plan–Evaluation (HOPE) and Assessment–Subjective–Objective–Plan (ASOP). Additional information and rationale can be found in [28–30].

There are several extensions of the SOAP format. In the original SOAP format, the source describes actions that will be taken in the treatment plan to resolve a problem given its occurrence making no distinction for hypothetical or conditional occurrences (e.g., if pain worsens, return to facility). In many cases, the plan may be executed at the time of care. Other extensions of the SOAP format, SOAPIE and SOAPIER, address this particular type of statement where (I) is an *implementation* or *intervention* describing actions taken or already executed at the current visit, (E) provides an *evaluation* of the patient's response to the given treatment and (R) describes any *revision* or *reassessment* of either the original or any subsequent plans [25]. Finally, the APIE and ADPIE formats condense the subjective and objective data into (A) *assessment* information for deriving a (D) *diagnosis* and then uses the last portion of the acronym (PIE) to articulate the *plan, implementation and evaluation* respectively [25,26].

2.3.2. Application of SOAP format

Today, medical students are taught the SOAP framework as a method to aid clinical decision-making and facilitate communication between clinical staff members. For instance, a SOAP note format is used to test fourth year medical student's communication skills through patient-based examinations on the United States Medical Licensing Examination Step 2 Clinical Skills Exam [31,32]. This is an important skill set that demonstrates the student's ability to apply medical knowledge in a clinical scenario under supervision. Once a SOAP note is documented, it becomes part of the official medical record used for clinical, financial and legal accountability. The SOAP note can serve as documentation communicating pertinent clinical information between members of the medical team, demonstrating that care was ‘‘reasonable and medically necessary'' to regulatory (CMS [33]) institutions and providing evidence of facts that might be subpoenaed by attorneys from legal (Health care legal team) agencies. Profession-specific guides and instructions for writing SOAP notes can be found in [24,25,27].

The goal of this empirical study was to develop and evaluate an automatic SOAP classifier that classifies sentences from emergency department reports. To fulfill this goal, we propose two objectives: (1) determine how well the SOAP framework applies to emergency department reports and (2) determine the types of features that support successful automated SOAP classification.

3. Methods

We conducted a quantitative study to develop a reliably annotated reference standard with the SOAP framework and determine what combinations of feature types are necessary for building an accurate SOAP classifier.

3.1. Developing the annotation schema

We constructed SOAP class definitions through both literature review and a pilot annotation study. For our purposes, we defined ***subjective*** as *background or historical information relevant to understanding the patient's current or future clinical state* and ***objective*** as *observable, measurable and quantifiable information*. We did not instruct annotators to use the source of the information, patient or care provider, as a major source for their determination.

We defined **assessment** as *expressions of a diagnosis, impression or differential diagnosis* and **plans** as *any reporting of planned or implemented treatment actions, education or follow-up procedures*. We conducted an initial pilot study on 10 emergency department reports ($n = 734$ sentences) not used in this study. From the pilot study, we clarified our definitions based on annotator feedback and agreement. Table 1 presents our final annotation schema for SOAP classes and definition for this follow up study.

3.2. Dataset

This study was approved by the University of Pittsburgh Institute Review Board (IRB). We obtained a dataset comprised of 50 emergency department (ED) reports from the University of Pittsburgh Medical Center (UPMC) aggregated from visits occurring from December 1990 to September 2003. This subset was randomly selected from a dataset described in [34]. The data were obtained from the MARS repository, a database that stores a variety of transcribed report types in which each report contains diagnosis codes and chief complaints [35].

3.2.1. Process for developing the reference standard

Two annotators, A1 and A2, each with over 10 years of experience in the clinical setting, annotated sentences from the emergency department reports. A1 is a Registered Health Information Administrator (RHIA) with over 15 years of work experience in the Department of Defense healthcare system; A2 is a licensed Registered Nurse with 10 years experience in various inpatient and ambulatory settings. The annotators were provided a 13-page instruction guide and annotated each sentence in the report with all SOAP classes that applied. Agreement was evaluated after the first five reports and again after the 25th report. At that point, we found agreement was consistently sufficient (kappa coefficient above 0.70 for all classes) for only the second annotator to annotate the remaining 25 reports. Disagreements in the first 25 reports were settled by randomly selecting one of the annotator's answers for all classes. The annotations were collected with a web-based annotation tool built using the Django infrastructure written in Python [36].

3.2.2. Data generation and processing

We generated a binary dataset from the manually annotated data for training and testing a classifier for each SOAP class. Given each SOAP class, every sentence was labeled as a positive or negative instance. We used the decision rule that if a sentence was annotated with a particular SOAP label, that sentence is a positive instance for that class; otherwise it is a negative instance for that

Table 1
SOAP classes and definitions.

SOAP class	Definition
Subjective	*Background or historical information that may be relevant to understanding the patient's current or future clinical state such as description of events leading from the last encounter to the current visit, pertinent past and family histories, social habits placing the patient at risk for disease, current medications used to manage existing conditions and known allergies*
Objective	*Observable, measurable or quantifiable data obtained from past records, physical examinations, tests, procedures, screenings and other diagnostic techniques*
Assessment	*Possible diagnosis including reported differentials and impressions by the dictating physician or clinical staff treating the patient*
Plan	*Completed or follow-up care plans, treatment actions, education*
Not applicable	*Irreconcilable statements that do not apply to the previous classes*

class. For example, sentences annotated as *plan* are positive instances for the *plan* class and all sentences not annotated as *plan* are negative instances. All *not applicable* sentences were labeled as negative instances for each class.

3.3. Experiments with classifiers

We investigated the following baselines, supervised algorithms and feature groups for creating and evaluating automated SOAP classifiers.

3.3.1. Baselines

To determine the complexity of the task, we initially developed simple baseline classifiers. The first baseline assigned the target class for every sentence in the reference standard (i.e., the class *objective* for the *objective* classifier, *etc.*). The second baseline assigned the majority class to every sentence. The third baseline used a conditional probability distribution to identify the most likely SOAP class for each section in the report. For each sentence, this classifier assigned the most likely SOAP class with the highest conditional probability, e.g., if the "disposition" section type was most likely to be assigned the *plan* class, all sentences in the "disposition" section were classified as *plan*. Sections were tagged using SecTag, an automated section tagger [18], and conditional probabilities were calculated using the Natural Language Tool Kit (NLTK) and the pilot dataset. Specifically, for each sentence, this classifier assigned the SOAP class with the highest posterior probability, e.g., if the "disposition" section type was more likely to be assigned as a *plan* class in the pilot set, all sentences in the "disposition" section in the test set were classified as *plan*. Table 2 contains examples of section header types correlated to SOAP classes.

3.3.2. Supervised classifier and experimental design

We created SOAP classifiers using a variety of feature groups and support vector machines, respectively.

3.3.2.1. Feature groups. We included a variety of features, including many designed to collapse similar features into a smaller set of values to reduce the feature space.

(1) Lexical: Lexical features comprise tokens found in the report. We used the natural language toolkit (NLTK) to identify all **unigrams** and **bigrams**. In "The patient has a history of stroke" the full set of lexical features include ⟨s⟩, *The, patient, has, a, history, of, stroke*, .,⟨/s⟩, ⟨s⟩ *The, The patient, patient has, has a, a history, history of, of stroke, stroke* ., .⟨/s⟩, where ⟨s⟩ and ⟨/s⟩ indicate the start and end of the sentence, respectively.

(2) Syntactic: Syntactic features consist of Penn Treebank tags [37] encoded by the Stanford part of speech tagger (09/28/2009) [38] and corrected for common tagging errors that occur in clinical narratives using seven rules that were learned by applying Brill's transformation-based tagger to a previous set of clinical reports [39]. For example, one of the rules states that if a token with the tag "CD" is followed by the token ".", change the tag "CD" to the tag "LS", indicating that the number is part of a numbered list. We identified the **part of speech** and **word/part of speech pair** (word/POS) for each lexical feature as a crude attempt at word sense disambiguation. For instance, "discharge" (NN) often indicates a clinical finding, whereas "discharge" (VB) indicates being released from the hospital.

For every verb phrase in the sentence, we encoded the **tense** of the first verb in each verb phrase as *past, present* or *future*. For example, we classified "She had developed a severe cough" as *past*, and "she will return if she develops a severe cough" as *future* and *present*, respectively.

(3) Semantic: We used the Unified Medical Language System (UMLS) Metathesaurus (version.2.4.C release) courtesy of the National Library of Medicine to tag the **semantic type** and **cui** (concept unique identifier) for each token in the sentence found in the UMLS [40]. For example, in the phrase "Lungs are clear", "Lungs" maps to the semantic type *Body Part, Organ, or Organ Component* and CUI: C0024109, and "clear" maps to semantic type *Idea or Concept* and CUI: C1550016. We also captured the **position of each semantic type** in the sentence as *Beginning, Middle*, or *End*, based on character counts within the sentence. We applied a feature reduction strategy [41] to encode whether a **digit type** was being used as a *date, list, anatomical location, medication, result* or *age*. We used simple regular expressions and heuristics to assign the digit type. For example, "1. aspirin" – *list*:, "cranial nerves II through XII are grossly intact" – *anatomic location*, "20 mg" – *medication*, and "Temp 98.6" – *result*.

The emergency department reports were de-identified according to the HIPAA criteria by DE-ID software (version 5.10). We used the **de-id tags** as features representing patient-sensitive or service facility information: *name, date, device-id* or *institution*.

We identified **state of mind** terms as shallow predictors of mental postulation suggestive of medical decision making and **hedge terms** from [42] suggestive of uncertainty and speculation. For example, in "I think he has viral meningitis," "think" was encoded as a **state of mind** term. Similarly, in "She likely has the flu," "likely" was encoded as a **hedge term**.

Finally, we included trigger terms applied by the ConText algorithm, which indicate that a condition in the sentence is *historical* (e.g., "history of"), *conditional* (e.g., "if"), *absent* (e.g., "denies"), or *experienced by someone other than the patient* (e.g., "family history") [43].

(4) Contextual: We defined the contextual information about the sentence with respect to the structure of the clinical narrative. We used the SecTag tagger to identify the **section type** for each sentence found in the report [18]. For example, "Cardiovascular: The patient has chest pain" maps to a section type *cardiovascular_review*. SecTag defines 16,036 possible section tags.

Because emergency department report structure may follow chronological ordering similar to ideal progress notes (i.e., Subjective < Objective < Assessment < Plan), we included a feature encoding the **position** of the sentence in the report in *quartiles*. We also included **length of the sentence** in *number of tokens*. For instance, "Chief Complaint: headache" is in the *1st quartile* of the report and has *length of 6* including sentence start and end markers.

(5) Heuristic: We developed an unsupervised method for mining **high-precision terms** from a corpus of de-identified emergency department reports (200,000 sentences from 3577 reports) from

Table 2
Example sections probabilistically associated with SOAP classes.

SOAP class	SecTag section header types
Subjective	*allergies_and_adverse_reactions, back_review, chief_complaint, family_and_social_history, family_medical_history, history_present_illness, hospital_course, medications, past_medical_history, past_personal_and_social_history, review_of_systems, risk_factors, substance_use, tobacco_use*
Objective	*abdominal_exam, chest_exam, counts, derm_exam, extremity_exam, general_exam, genitourinary_exam, head_neck_exam, heart_rate, heent_course, hematology_exam, laboratory_and_radiology_data, laboratory_data, pelvis_exam,*
Assessment	*admission_diagnosis, diagnoses, discharge_condition, discharge_diagnosis*
Plan	*discharge_medications, disposition_plan, ear_nose_throat_exam, follow_up*

Table 3
Features used to train supervised SOAP classifiers. Values for all features were binary: 1 if the feature was present in the sentence and 0 if the feature was not.

Feature group	Featavure type	Description	*Features*
Lexical	Unigrams	Each token in the sentence	*chest, pain patient, treated, etc*
	Bigrams	Each contiguous pair of unigrams in the sentence	*chest_pain, secondary_to, etc*
Syntactic	Part of speech (pos)	Penn Treebank part-of-speech tags present in the sentence	*VBN, NN, MD, etc*
	Word/POS pair	Each token in the sentence conjoined with its part of speech	*discharge_NN, treated_VBD, etc*
	Verb tense	Tense of first verb of each verb phrase in the sentence	*Past, Present, Future*
Semantic	UMLS semantic type	Semantic types in the sentence	*Idea or Concept, Sign or Symptom, etc*
	UMLS CUI	Concept unique identifiers in the sentence	*C0205168, C0024109, C1550016, etc*
	UMLS semantic type sentence position	Semantic type conjoined with its position in sentence (beginning, middle, end)	*Idea or Concept_End, Sign or Symptom_beginning, etc*
	Digit type	Function of the digits in the sentence	*LIST, RESULT, DATE, ANATOMICAL LOCATION, MEDICATION, AGE*
	De-identification tags	Type of de-identification tag	*NAME, AGE, DEVICE-ID, etc*
	ConText lexicon	Trigger terms used by ConText	*history, if, denies, mother, etc*
	State of Mind lexicon	Mental postulation terms	*think, believe, know, etc*
	Hedge lexicon	Certainty and speculation terms	*likely, probable, might, possibility, etc*
Contextual	SecTag section type	Report section in which the sentence resides	*vital_signs, admission_diagnosis, etc*
	Quartile position	Quartile position of sentence in report	*1st, 2nd, 3rd or 4th*
	Sentence length	Length measured by tokens	*length_5, length_7, etc. (ranging from 3 to 79)*
Heuristic	Heuristic lexicon	Unigrams, bigrams and phrases thought to be predictive of a SOAP class	*will_need, positive, presents_with, my_impression, etc*

the University of Pittsburgh NLP Repository [44,45]. We used an initial seed set of 5–6 terms to predict the SOAP class for each sentence by assuming all sentences that contained the seed terms belong to that class. From these tagged sentences, we used a simple conditional probability to learn new terms as good predictors for a SOAP class. For example, if "alcohol" is a *subjective* seed and tags the sentence "patient drinks one glass of alcohol a day", the conditional probability may learn "drinks" as a correlated term for *subjective*. Additionally, we conducted an error analysis on our pilot data to identify phrases we thought would be indicative of each class.

For every sentence in the corpus, we created a vector of features with binary values to indicate whether or not that feature was present in the sentence. Features representing words or classes from the text (e.g., unigrams or UMLS semantic type) were generated from the pilot set so a feature not present in the pilot set was not applied to this dataset. Table 3 describes the features and provides examples.

3.3.2.2. Support vector machines. First developed in 1995 [46], the support vector machine (SVM) determines a hypothesis or model that linearly separates two classes in a high dimensional space [46,47]. We chose to build SOAP classifiers using SVMs, because SVMs (1) handle a large number of features while maintaining high performance, (2) reduce the likelihood of overfitting by using support vectors for classification, and (3) tolerate sparse data vectors that may be produced encoding a high number of features explained in [47] and used in similar studies such as [1,41,47,48]. We used the SVM implementation in Weka (version 3.6), a standard machine learning software package, with a linear function, default settings and 10-fold cross validation [49]. In our pilot study, we experimented with naïve Bayes and Decision Tree classifiers but found that they produced inferior results compared to SVM. We suspect that naïve Bayes performance degraded as we increased the number of features due to the violation of the conditional independence assumption among features, and the Decision Tree performed poorly since the features added did not provide sufficient information gain.

All features were automatically generated with programs we implemented in Python version 2.5. For each sentence, we encoded the feature value as "1" if the feature was found in the sentence and "0" otherwise. Each set of similar features was mapped to one of five feature groups: *lexical, syntactic, semantic, contextual* or *heuristic* (details in 3.3.2.2). For each SOAP class, we trained and tested two SVMs. The first was trained on all features. The second was trained on only those features that were included by chi-square feature selection with a significance threshold of $p < 0.05$. We compared the output of all classifiers against the manual reference standard to address four questions: (1) *How well does a classifier perform when trained on all feature groups?* (2) *How well does a classifier perform when trained on a subset of features selected through a feature selection algorithm?* (3) *How much does each feature group contribute to performance on the classification task?* (4) *Which feature group is most informative to the classification task as a whole?* To answer question (1) we trained an inclusive classifier using all feature groups, (2) we trained a classifier using a subset of features selected with a feature selection algorithm, (3) we trained classifiers using each feature group individually, and (4) we trained classifiers by leaving out one feature group at a time (ablation study). Fig. 1 depicts each of these arms.

3.4. Dataset evaluation metrics

We measured inter-annotator agreement for the expert annotations and predictive performance of the SOAP classifiers when compared against the expert-generated reference standard.

3.4.1. Inter-annotator agreement

Out first objective was to determine how well the SOAP framework applies to emergency department reports. We determined the prevalence of each class in our corpus and measured the proportion of sentences not classified into a SOAP class.

To determine whether annotators can annotate the SOAP model with high level of agreement, we computed inter-annotator agreement metrics described in [50,51]. Low inter-annotator agreement can indicate that the instructions may not be clear, the annotation schema might not be intuitive, or the task may be too difficult to perform. We calculated observed agreement, along with positive and negative specific agreement. To account for chance agreement, we computed kappa metrics according to [50–53]. We considered a kappa coefficient greater than 0.70 as acceptable agreement and as an indication of the appropriateness of the SOAP model for the

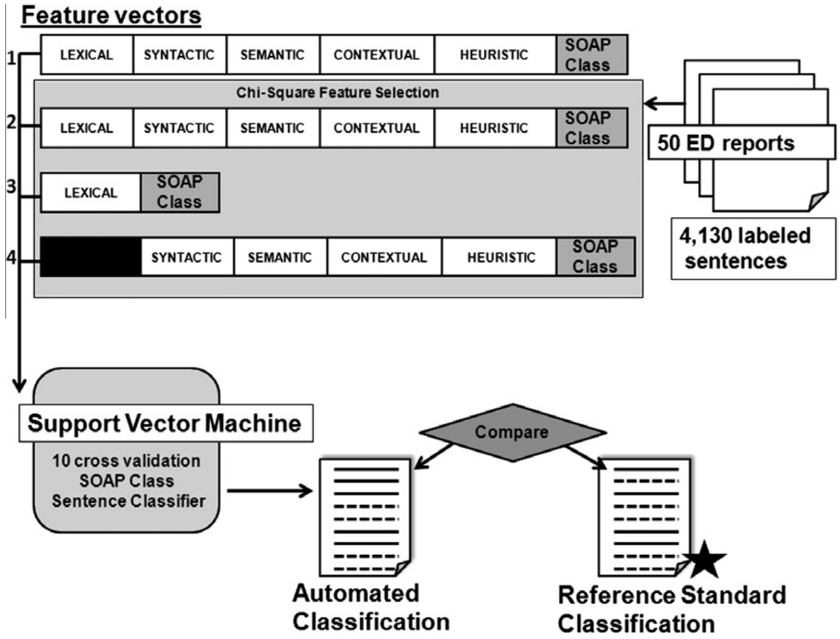

Fig. 1. Experimental design for determining performance with (1) all feature groups, (2) all features included through feature selection, (3) each feature group individually, and (4) each feature group individually held out.

Table 4
Inter-annotator agreement metrics.

Agreement metrics	Formulas
Observed	$A_o = (TP + TN)/(TP + TN + FP + FN)$
Positive specific	$Pos = 2(TP)/(2 * TP + FP + FN)$
Negative specific	$Neg = 2(TN)/(2 * TN + FP + FN)$
Cohen's kappa (chance-corrected)	k (chance) = $A_o - A_e/1 - A_e$
Kappa (prevalence-corrected)	k (prevalence) = $2(A_o) - 1$

Table 6
Performance metrics and corresponding formulas.

Performance metric	Formula
Accuracy	$Acc = TP + TN/TP + TN + FP + FN$
Recall (sensitivity)	$Rec = TP/TP + FN$
Precision (positive predictive value)	$Prec = TP/TP + FP$
F_1 score	$F_1 = 2(Prec * Rec)/(Prec + Rec)$
McNemar's Test w/Yates correction	$X^2 = (\lvert FP - FN\rvert - 0.5)^2/(FP + FN)$

corpus. See Table 4 for a breakdown of each agreement metric and corresponding formula.

3.4.2. Evaluating the classifiers

To evaluate how well each classifier identified each SOAP class, we used standard evaluation metrics: accuracy, recall, and precision. We computed the F_1 score, which represents the harmonic mean between recall and precision and used the F_1 score to select the best performing classifier. We used McNemar's test (X^2) to evaluate whether the classifier errors were statistically significantly different for classifiers trained on all feature groups and classifiers trained after feature selection. We applied Yates correction (0.50) when one cell in the contingency table was less than or equal to 5 [54]. Table 5 defines our variables and table 6 defines the formula for each performance metric.

4. Results

We measured inter-annotator agreement of expert annotators applying the SOAP model to ED reports and developed SOAP classifiers using a diverse number of features. We observed the following results.

Table 5
Definitions used to calculate evaluation metrics for the presence or absence of a particular SOAP class.

	Reference standard	Automated classification
True positive (TP)	Present	Present
True negative (TN)	Absent	Absent
False positive (FP)	Absent	Present
False negative (FN)	Present	Absent

4.1. Reference standard characteristics and annotation study

Our dataset of 50 reports was comprised 4130 sentences in which the number of sentences per document ranged from 32 to 198, with an average of 82.6 sentences per document. Prevalence and frequency of SOAP classes in the 4130 sentences was as follows: 35.5% (*subjective; n = 1468*), 44.0% (*objective; n = 1818*), 5.5% (*assessment; n = 227*), 11.3% (*plan; n = 465*), and 8.1% (*not applicable; n = 335*). We observed that 3956 (95.8%) of sentences can be classified by a single SOAP class, 165 (4.0%) sentences have two classes, and 9 (0.02%) have three classes. Inter-annotator agreement for all classes exceeded the threshold for adequate agreement (0.70), as shown in Table 7. The most prevalent classes, *subjective* and *objective*, demonstrated greater than 0.90 agreement across all agreement metrics. Agreement was lowest for *assessment* with a Cohen's kappa of 0.76; however, once corrected for prevalence, the kappa value increased to 0.940.

4.2. Predictive performance of automated SOAP classification

We generated 32,215 features, the majority of which were unigrams, bigrams and POS/word pairs. As shown in Table 8, automated feature selection reduced the number of features from 32,215 to between 1911 (for *assessment*) and 5549 (for *plan*). ConText lexicon, UMLS semantic type, and UMLS semantic type sentence position were completely removed during feature selection,

Table 7
Inter-annotator agreement on the corpus.

SOAP category	Observed agreement	Positive specific agreement	Negative specific agreement	Cohen's kappa	Kappa corrected for prevalence
Subjective	0.971	0.958	0.978	0.936	0.940
Objective	0.954	0.946	0.961	0.907	0.910
Assessment	0.970	0.776	0.984	0.760	0.940
Plan	0.959	0.826	0.976	0.802	0.920

indicating that those features do not provide useful or independent information for classification. After feature selection, the lexical features (unigrams and bigrams), word/POS pairs, UMLS CUIs and SecTag section types comprise the majority of features across all SOAP classes.

Table 9 shows predictive performance of all SOAP classifiers. Overall, most supervised classifiers outperformed the baseline classifiers. As expected, the Positive class baseline did not have adequate precision, resulting in poor F_1 scores for the less prevalent classes, *assessment* (11.0) and *plan* (22.5). The Majority class baseline did not predict the SOAP class but reflected the imbalanced class distribution in the dataset. The Section classifier performed quite well with high F_1 scores for *subjective* (88.2) and *objective* (70.2); however, it produced moderate performance for the less prevalent classes of *assessment* (54.4) and *plan* (70.2). The Section classifier performed with low recall on *assessment* (50.0) and *plan* (20.6) classes.

The SOAP classifier without feature selection (w/o FS) outperformed the Section classifier baseline by increasing the points of F_1 scores for all classes – 6.2 (*subjective*), 23.2 (*objective*), 7.7 (*assessment*) and 42.6 (*plan*). The improved F_1 scores can be explained by increased coverage for most classes with recall gains of 5.3 points (*subjective*), 34.8 points (*objective*) and 47.0 points (*plan*). We observed these gains at no expense of precision but instead with modest to substantial point increases of 7.4 (*subjective*), 4.8 (*objective*), 29.6 (*assessment*) and 3.8 (*plan*).

We applied feature selection (w/FS) to reduce the feature space and determine if we could further improve the F_1 scores. Feature selection improved performance for most classes, showing gains ranging from 1.1 to 13.6 points with the exception of the *subjective*

Table 8
Feature type and counts for each class before and after feature selection (FS).

Feature type	Before FS	Subjective	Objective	Assessment	Plan
Unigrams	3950	520	559	218	708
Bigrams	19,478	987	1026	1144	3453
Part of speech (POS)	40	15	22	14	24
Word/POS pair	5183	532	578	349	858
Verb tense	3	3	3	0	1
UMLS semantic type	192	0	0	0	0
UMLS CUI	2769	328	344	152	443
UMLS semantic type sentence position	159	0	0	0	0
Digit type	6	2	1	0	0
De-identification tags	4	3	1	0	1
ConText lexicon	117	0	0	0	0
State of mind lexicon	26	2	3	1	1
Hedge lexicon	51	5	5	9	4
SecTag section type	113	55	56	17	26
Quartile position	4	3	3	3	3
Sentence length	72	5	5	2	20
Heuristic lexicon	48	13	7	2	7
Total features	32,215	2473	2613	1911	5549

Table 9
SOAP Classifiers including baselines (positive and majority class and section), all feature groups with and without feature selection (FS), each individual feature group (Lex = lexical, Syn = syntactic, Sem = semantic, Con = contextual, Heur = heuristic) and ablation arms (sans or leave-one-group-out).

Classifiers	Subjective				Objective				Assessment				Plan			
	Acc	F_1	Prec	Rec	Acc	F_1	Prec	Rec	Acc	F_1	Prec	Rec	Acc	F_1	Prec	Rec
Positive	**35.5**	**70.8**	**35.5**	**100**	**44.0**	**87.6**	**44.0**	**100**	**5.5**	**11.0**	**5.5**	**100**	**11.3**	**22.5**	**11.3**	**100**
Majority	**65.5**	**0**	**0**	**0**	**56.0**	**0**	**0**	**0**	**94.5**	**0**	**0**	**0**	**88.7**	**0**	**0**	**0**
Section	**91.8**	**88.2**	**90.2**	**86.2**	**78.4**	**70.2**	**89.1**	**58.0**	**95.4**	**54.4**	**60.1**	**50.0**	**90.6**	**33.0**	**82.0**	**20.6**
w/o FS	96.2	94.4	97.6	91.5	94.2	93.4	93.9	92.8	96.8	62.1	89.7	47.5	95.2	75.6	85.8	67.6
w/FS	95.7	93.9	94.7	93.1	95.2	94.5	94.5	94.5	97.8	75.7	95.9	62.6	95.5	77.0	90.7	66.9
Lex only	**93.5**	**90.6**	**92.9**	**88.5**	93.5	92.6	92.6	92.6	**97.3**	**69.3**	**94.7**	**54.6**	**94.9**	**73.0**	**91.3**	**60.9**
Syn only	**91.5**	**87.7**	**91.0**	**84.5**	93.0	92.0	92.8	91.2	**96.7**	**60.1**	**92.7**	**44.5**	94.8	73.4	85.9	64.1
Sem only	**88.4**	**82.0**	**91.0**	**74.7**	**89.5**	**87.1**	**95.0**	**80.4**	**96.5**	**58.0**	**83.5**	**44.5**	**93.4**	**65.0**	**81.3**	**54.2**
Con only	**91.5**	**87.5**	**91.4**	**83.9**	**86.2**	**84.7**	**83.3**	**86.1**	**95.4**	**40.0**	**71.6**	**27.8**	**91.0**	**46.7**	**70.7**	**34.8**
Heur only	**68.9**	**31.4**	**73.0**	**20.0**	**55.9**	**1.4**	**44.8**	**.70**	**94.5**	**0**	**0**	**0**	**88.7**	**0**	**0**	**0**
Sans Lex	96.0	94.4	95.5	93.3	94.8	94.1	94.3	93.8	**97.8**	**76.6**	**92.0**	**65.6**	**95.7**	**78.9**	**88.5**	**71.2**
Sans Syn	95.9	94.1	94.9	93.4	95.4	94.8	94.8	94.8	97.5	72.7	92.5	59.9	95.6	77.5	92.0	66.9
Sans Sem	95.9	94.1	95.2	93.1	95.1	94.4	94.5	94.4	97.7	74.1	95.8	60.4	95.6	77.2	91.7	66.7
Sans Con	**94.1**	**91.5**	**94.2**	**89.0**	94.0	93.2	93.4	93.1	**97.6**	**72.5**	**96.4**	**58.1**	95.0	74.0	88.9	63.4
Sans Heur	95.8	94.0	94.8	93.2	95.0	94.4	94.4	94.4	97.8	76.2	95.4	63.4	95.6	77.2	91.7	66.7

Table 10
Feature values with the 15 highest weights for each SOAP class.

Subjective	Objective	Assessment	Plan
past_medical_history	rectal_exam	discharge_condition	date_transcribed
history_present_illness	cardiovascular_exam	admission_diagnosis	reviewed_VBN
allergies_and_adverse_rxns	heent_exam	discharge_diagnosis	discharge_VB
review_of_systems	abdominal_exam	C0042029	follow_VB
ear_review	extremity_exam	C0851827	"reviewed with"
cardiovascular_review	general_exam	weakness_NNP	"admitted"
gastrointestinal_review	neurological_exam	assessment_NNP	"the plan"
neurologic_review	C0015385	GI_RB	"this plan"
medications	C0205307	assessment_NN	"⟨s⟩ follow"
C1301808	C0007012	dehydrated_VBN	"evaluated by"
C0027497	elevated_VBN	noninsulin-dependent_JJ	"a lumbar"
C0030450	CO2_NNP	"confusion."	"puncture without"
C0332272	not_RB	":confusion"	"examination findings"
": negative"	he_states	"to micu"	"a bit"
": no"	"sent."	"micu for"	"⟨s⟩ I"

class, which dropped by 0.5 points. In evaluating how well each feature group performed individually, we found that no single feature group individually produced greater F_1 scores than the SOAP classifiers w/FS. Finally, we assessed how informative each feature group was to SOAP class prediction using an ablation study design. For each class, we observed a reduction of F_1 scores by removing the contextual feature group, which was largely due to decreases in recall without the contextual features. This finding indicates that contextual features are important to SOAP classification.

From the feature selection algorithm, we identified the most informative features for predicting each of the four SOAP classes. Section categories, CUIs, unigrams, bigrams and word/POS pairs were among the feature values with the highest weights (Table 10).

Sentences with one SOAP class were more accurately predicted than sentences with more than one SOAP class (Table 11), suggesting that the classifiers may require more features to distinguish the presence of the class when more than one SOAP class is present.

5. Discussion

The objectives of this study were to (1) assess the applicability of the SOAP model for ED reports and (2) determine which features contribute to accurate SOAP classification.

5.1. Applicability of the SOAP model for ED reports

The SOAP model applied to 3836 (92.9%) sentences in our dataset. All sentences that were not assigned a SOAP class by annotators either served administrative purposes, such as "Signed by: **NAME[AAA XXX GGG], MD," or were incorrectly segmented sentences, such as a section heading like "PHYSICAL EXAM:" segmented as a sentence. There were several sentences with more than one class assigned (3.99%). In rare cases, multiple class assignment was due to incorrect sentence segmentation in which two sentences were segmented as one. Most sentences with multiple SOAP classifications represented descriptions of clinical reasoning relating, for example, an *objective* measurement to a *plan* or a *plan* to an *assessment*. For instance, "She will be discharged in good condition with impression of viral illness" consists of both *plan* (patient will be discharged) and *assessment* (impression that she has a viral illness).

Table 11
Performance for sentences (1–3 classes) for the dataset using all feature groups w/FS.

	Accuracy	F_1 score	Recall	Precision
w/FS (1 class) n = 3956				
Subjective	96.2	94.7	94.2	95.3
Objective	96.0	95.4	96.2	94.6
Assessment	98.9	84.5	75.5	95.9
Plan	96.9	78.3	69.8	89.2
w/FS (2 classes) n = 165				
Subjective	86.1	54.9	50.0	60.9
Objective	80.0	77.2	67.5	90.3
Assessment	72.7	50.5	34.3	95.8
Plan	70.9	73.9	60.2	95.8
w/FS (3 classes) n = 9				
Subjective	44.4	0.0	0.0	0.0
Objective	11.1	20.0	11.1	100.0
Assessment	66.7	57.1	40.0	100.0
Plan	44.4	54.5	37.5	100.0

Annotators showed high agreement on the SOAP annotation task. The coverage of SOAP classes and high agreement for expert annotation suggests that the SOAP framework is applicable to ED reports and that the annotation schema for SOAP classes was well defined. Performing the pilot study was helpful in evolving a schema for human experts. We also suspect that giving annotators flexibility of assigning all classes that apply to a single sentence was important for eliciting good agreement. Disagreements between annotators occurred most often when a statement was an *assessment*. This was consistent with our finding during the pilot study and was often a point of disagreement during training. For example, one annotator consistently labeled some sentences containing a disease name as an *assessment* even when the context indicated a *plan*, like "UTI" in "He was given printed instructions about UTI, Pyridium, and ciprofloxacin".

5.2. Performance of Automated SOAP Classifiers

Prevalence of SOAP classifications in our dataset varied from 6% for *assessment* to 44% for *objective* classes. As expected, performance of the SVM classifier was higher for more prevalent classes, with F_1-scores of 0.62 (*assessment*), 0.76 (*plan*), 0.94 (*subjective*), and 0.93 (*objective*). Precision was higher than recall for all classes (the lowest precision score was 0.86 for *plan*), suggesting that false positive classifications were less of a problem than false negative classification and that more training data could further improve performance.

Feature selection tended to improve classification performance—especially for the two less prevalent classes, which showed 1.4 point (*plan*) and 13.6 point (*assessment*) increases. From 32,215 original features, the number of features was reduced by a range of 82.7% (*plan*) to 94.1% (*assessment*), indicating that most features were not needed for accurate classification or that features we included had overlapping information. For some

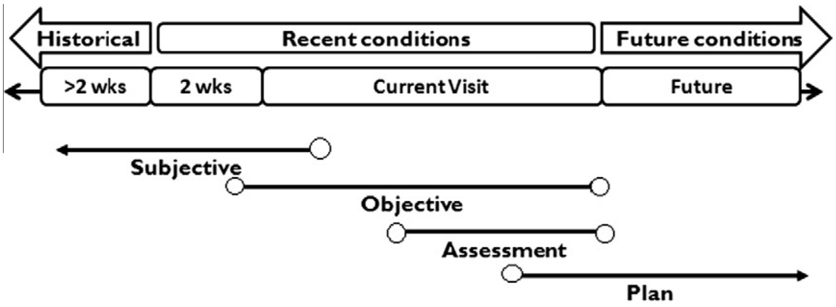

Fig. 2. Temporal progression of conditions mentioned in clinical text.

features, the number of values was reduced, including unigrams, bigrams, UMLS CUIs, section tags, and part-of-speech tags (see Table 8). For other features, all values for that feature were eliminated, including UMLS semantic type and its position in the sentence and the ConText lexicon. Eliminated features all belonged to the semantic feature type, but many were probably too broad to be discriminatory. For example, concepts with the UMLS semantic type "Body Part, Organ, or Organ Component" can occur in a description of review of systems, which is *subjective*, and in a description of a physical exam, which is *objective,* and therefore may not distinguish between the two classes. Semantic features that were not eliminated conveyed more specific information about the concept, such as the UMLS CUI C0015385 for limbs, or about the reasoning process of the physician, such as hedge terms or words indicating state of mind. Features with the highest positive weights included section headings (discharge_diagnosis for *assessment* and abdominal_exam for *objective*), and predictive unigrams and bigrams (": negative" for *subjective* and "admitted" for *plan*), and UMLS CUIs (C0205307 (normal) for *objective* and C0851827 (diabetes mellitus) for *assessment*).

We evaluated the contribution of different types of features to classification accuracy. Assigning a SOAP class based only on the section in which a sentence was found was less accurate than a classifier using all features, but was quite accurate for *subjective* (F_1 score 0.88) and *objective* (F_1 score 0.70) classes. Recall was especially low with the section classifier for all classes but *subjective* (0.86 *subjective*, 0.58 *objective*, 0.50 *assessment*, 0.21 *plan*). This finding is consistent with our experience in other classification tasks [55], showing that section is a critical factor in interpreting the context of a clinical condition but is not reliable enough to be the sole factor in classification. Good performance of the section-based SOAP classifier suggests that our map from sections to SOAP classes was effective and that SecTag performed well at automatically tagging sections. For example, the ability to distinguish the section *medications* from *discharge_medications* was critical to accurate assignment of *subjective* and *plan* classes. Reports from other institutions and other report types may be less amenable to automated section tagging.

No classifier trained on an individual feature group produced an F_1 score better than the classifier using all feature groups w/FS. However, performances of the syntactic feature group on *objective* and *plan* classification and of lexical features on *objective* classification were not statistically different from performance when using all groups. When we removed individual feature groups in the ablation studies, performance generally did not decrease significantly. Comparable performance may be due to overlap in feature values. For instance, the presence of lexical features such as unigrams and bigrams may provide enough information to discriminate the class when other features like state of mind and hedge terms were held out. Removing contextual features, such as sentence length and quartile position in a report, significantly decreased F_1 scores for classifying *subjective* and *assessment* classes suggests that location within the structure of a document is meaningful. One interesting and unexpected finding was that removing lexical features produced a higher F_1 score for classifying *assessment* and *plan* sentences. It may be that not relying on the words in the text can result in better performance when there is sparse training data.

During the pilot study, we performed a detailed error analysis on less prevalent classes and identified phrases and terms we thought would improve classification performance. We included the hand-crafted phrases as the heuristic feature group in this study and found that they did not provide any useful knowledge for predicting *assessment* and *plan*. This result may be due to the fact that the hand-crafted phrases are not very frequent, or it may be that the phrases are an indication of overfitting to our pilot data.

We reviewed the most heavily-weighted features for each of the target SOAP classes. For the *subjective* class, the most predictive features included sections that describe past and recent history, subsections of the review of systems, as well as CUIs attributed to "Signs and Symptoms", "Qualitative Concept" and "Geographic Locations." We would expect this, since physicians often describe the symptoms of the patient in terms of quality, severity and onset. In contrast, sections attributed to physical examination and CUIs associated with "Body Part, Organ, or Organ Component", "Functional Group" and "Biologically Active Substance, Inorganic Chemical" were predictive of *objective* sentences, which is consistent with our intuition that physicians describe findings and observations for each of the body systems and describe results from diagnostic tests and laboratories. Many of the features most predictive of *assessment* included diagnosis sections and CUIs describing "Population Groups" and "Disease or Syndrome." For the *plan* class, section tags were not highly predictive. We suspect this can be explained by the fact that physicians tend not to adhere to document structure as strictly at the end of a report as they do in the initial portion of the report, i.e., *Plan* and *Assessment* tend to become condensed into the ED Course as reports of implemented treatments, medical decision making and potential plans for follow-up. We also found that the word sense of a unigram is important for determining if a statement is a *plan,* such as discharge_VB versus discharge_NN.

6. Limitations

We have identified several limitations in this study. We found no other annotation study in which researchers applied the SOAP framework or developed SOAP classifiers for clinical text, so it is difficult to compare our results to other studies. Furthermore, we assessed reports from only one institution, and prevalence of SOAP categories may differ across institutions. The limited sample size, 50 emergency department reports, had performance implications for less prevalent classes. We constrained our study to one report type and cannot make claims about the SOAP model for other report types. However, our experience evaluating attribute assignment [43,55] and characterizing anaphoric reference in clinical reports [56] has shown that report types like History and Physical Exams, Progress Notes and Discharge Summaries are similar to ED reports in structure and lexical distribution, and SOAP classification in these narrative reports may also be successful. We conducted the classification task at the sentence unit and enabled annotators to label multiple classes; however, some applications may require a smaller unit (clinical concept or phrase) for SOAP classification. Applying SOAP classification to an individual concept or phrase may require additional features or a different approach.

7. Future work

SOAP classification could be a useful feature in other types of classification. For instance, because the SOAP framework seems to embody temporal progression, a temporal classifier that incorporates SOAP assignment as a feature may help discriminate whether a condition is historical (occurs in the past), recent (occurs shortly before or during the current visit) or not particular (occurs in a hypothetical or conditional context) (Fig. 2 illustrates this point). We plan to use the SOAP framework as a coarse "back bone" in a more fine-grained discourse schema by extending the classes to capture more detail. For example, we could further classify *plan* sentences as "implemented," "in progress," or "scheduled" or use the SOAPIER extension to further differentiate this class. Similarly, we suspect we will need to expand the model to disambiguate *subjective* sentences that contain events and conditions from the distant past from those events and symptoms from the recent past (just before or during the current encounter).

We plan to expand the SOAP framework into a more detailed schema for modeling the discourse structure found in clinical reports. We will investigate the contribution of the discourse framework in the context of information extraction tasks, e.g., generation of problem lists.

8. Conclusion

We showed that the SOAP model can be annotated with high inter-annotator agreement and that it provides excellent coverage for sentences in emergency department reports. The diverse features we used resulted in accurate automated assignment of *subjective* and *objective* classes and of fair assignment of *assessment* and *plan* classes. There is a tradeoff between the cost of acquiring syntactic and semantic features and the modest improvement over lexical features. SOAP classification of sentences could be a useful feature in other NLP tasks and could help localize information in reports for use in visualization and assessment of clinical care.

Acknowledgments

This study was approved by the University of Pittsburgh IRB. We acknowledge and thank our funding sources: NLM Fellowship 5T15LM007059 and R01LM009427-01. We would like to thank our two anonymous reviewers for their suggestions and comments. We would like to extend appreciation to the annotators for their work. A warm thanks to Rebecca Hwa PhD, Pam Jordan PhD and Brian E. Chapman PhD for providing suggestions about approaching the task and guidance developing the annotation tool. We thank the University of Pittsburgh Clinical Research Services, University of Pittsburgh NLP Repository and open-source community for de-identifying our data, making de-identified data openly available and providing open source tools to further clinical research in informatics, respectively.

References

[1] Doan S, Conway M, Collier N. An empirical study of sections in classifying disease outbreak reports. In: Lazakidou A, editor. Annals of information systems: web-based applications in healthcare and biomedicine. Springer Science+Business Media, LLC; 2010. p. 47–58.
[2] Aronsky D, Haug P. Diagnosing community-acquired pneumonia with a bayesian network. In: Proceedings of AMIA symposium, 1998. p. 632–6.
[3] Hyun S, Johnson SB, Bakken S. Exploring the ability of natural language processing to extract data from nursing narratives. CIN: Comput, Inform, Nurs 2009;4:215–23.
[4] Minsuk L, Cimino J, Zhu HR, Sable C, Shanker V, Ely J, et al. Beyond Information retrieval – medical question answering. In: AMIA annual symposium proceedings, 2006. p. 496–73.
[5] Wang X, Chase H, Hripcsak G, Friedman C. Selecting information for electronic health records for knowledge acquisition. J Biomed Inform 2010;43:595–601.
[6] Weed L. Medical records. Medical education and patient care: the problem-oriented record as a basic tool. Year Book Medical Publishers: Press of Case Western Reserve University Cleveland; 1970.
[7] Solti I, Aaronson B, Fletcher G, Solti M, Gennari JH, Cooper M, et al. Building an automated problem list based on natural language processing: lessons learned in the early phase of development AMIA. Ann Symp Proc 2008:687–91.
[8] Meystre S, Haug PJ. Improving the sensitivity of the problem list in an intensive care unit by using natural language processing, AMIA. Ann Symp Proc 2006:554–8.
[9] Cao H, Chiang MF, Cimino JJ, Friedman C, Hripcsak G. Automatic summarization of patient discharge summaries to create problem lists using medical language processing, medinfo. Amsterdam: IOS Press; 2004. p. 1540.
[10] Meystre S, Haug PJ. Automation of a problem list using natural language processing. BMC Med Inform Decis Mak 2005:16.
[11] JCAHO. Standard IM 6.40. In: Smith IJ, editor. The joint commission guide to improving staff communication, 2005. p. 22.
[12] IOM. In the computer-based patient record: an essential technology for health care. In: Dick R, Steen E, Detmer D, editors. Washington (DC): National Academy Press; 1997.
[13] CMS. Eligible professional meaningful use core measures: measure 3 of 15, <http://www.cms.gov/EHRIncentivePrograms/Downloads/3MaintainProblemList.pdf>; 2010 [retrieved 2.08.11].
[14] AHIMA. Appendix A: definitions of problem lists from authoritative sources, <http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_036240.pdf>; 2011 [retrieved 8.08.11].
[15] Vleck TT Van, Wilcox A, Stetson P, Johnson SB, Elhaded N. Content and structure of clinical problem lists: a corpus analysis, AMIA. Ann Symp Proc 2008:753–7.
[16] Utiyama M, Isahara H. A statistical model for domain-independent text segmentation. In: Proceedings of the 39th annual meeting on association for computational linguistics, Stroudsburg (PA): Association for Computational Linguistics; 2001. p. 499–506.
[17] Apostolova E, Channin DS, Demner-Fushman D, Furst J, Lytinen S, Raicu D. Automatic segmentation of clinical texts. Conf Proc IEEE Eng Med Biol Soc 2009:5905–8.
[18] Denny JC, Spickard A, Johnson KB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. J Am Med Inform Assoc 2009:806–15.
[19] Cho PS, Taira RK, Kangarloo H. Automatic section segmentation of medical reports, AMIA. Ann Symp Proc 2003:155–9.
[20] Bhowmick PK, Basu A, Mitra P. Classifying emotions in news sentences: when machine classification meets human classification. Int J Comput Sci Eng 2010;2:98–108.
[21] Hachey B, Grover C. Sequence modelling for sentence classification in a legal summarisation system, 2005 ACM Symposium on Applied Computing; 2005. p. 292–6.
[22] Ruch P, Boyer C, Chichester C, Tbahriti I, Geissbuhler A, Fabry P, et al. Using argumentation to extract key sentences from biomedical abstracts. Int J Med Inform 2007;76:195–200.
[23] Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support evidence-based medicine. BMC Bioinform 2011;12:S5.
[24] Cameron S, Turtle-Song I. Learning to write case notes using SOAP format. J Couns Develop 2002:286–92.
[25] Kozier B, Erb G, Blais K, Johnson JY, Temple JS. Techniques in clinical nursing. New York: Addison-Wesley Nursing; 1993.
[26] Delegates Ho. Exhibit 7: competencies for entry into the profession of dental hygiene. J Dent Educ 2003:929–33.
[27] Borcherding S, Morreale MJ. The OTA's guide to writing SOAP notes. NJ: SLACK Incorporated; 2006.
[28] Bhopal J. Simple SOAP system. Br Med J 1981;283:892.
[29] Genisinger R, Fowler J. ASOP: a new method and tools for capturing a clinical encounter, AMIA. Ann Symp Proc 1995:142–6.
[30] Sorgente T, Fernandez EB, Larrondo Petrie MM. SOAP pattern for medical charts, 2008.
[31] McCauley M. The step 2 clinical skills exam: just another hurdle, American College of Physicians: Internal Medicine. Philadelphia: American College of Physicians; 2005.

[32] USMLE. USMLE Step 2 CS Content Description and General Information Booklet, FSMB and NBME, <http://www.usmle.org/Examinations/step2/step2cs_content.html>; 2011 [retrieved 2.08.11].

[33] Social Security Act, 1862 (a) (1) (A). Compilations of the social security laws. Baltimore; 2011.

[34] Chapman WW, Dowling JN, Wagner MM. Generating a reliable reference standard set for syndromic case classification. J Am Med Inform Assoc 2005;12:618–29.

[35] Yount RJ, Vries JK, Councill CD. The medical archival system: an information retrieval system based on parallel processing. Inf Process Manag 1991:379–89.

[36] Django. The web framework for perfectionists with deadlines, Django Software Foundation, <http://www.djangoproject.com>; 2009 [retrieved 6.08.09].

[37] Santorini B. Part-of-speech tagging guidelines for the penn treebank project. University of Pennsylvania: Department of Computer and Information Science; 1990.

[38] Toutanova K, Klein D, Manning CD, Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. HLT-NAACL 2003:252–9.

[39] Brill E. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. J Comput Linguist 1995;21:543–65.

[40] Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc 1998;5:1–11.

[41] Khoo A, Marom Y, Albrecht D. Experiments with sentence classification. In: Proceedings of the 2006 Australasian language technology workshop; 2006. p. 18–25.

[42] Mercer RE, Marco CD, Kroon FW. The frequency of hedge terms in citation contexts in scientific writing. Adv Artif Intell 2004:75–88.

[43] Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experience, and temporal status in clinical reports. J Biomed Inform 2009;42:839–51.

[44] Mowery D, Harkema H, Chapman BE, Hwa R, Wiebe J, Chapman WW. An automated SOAP classifier for emergency department reports. In: AMIA annual symposium proceedings, Washington (DC); 2010.

[45] University of Pittsburgh NLP Repository, <http://nlp.dbmi.pitt.edu/nlpfront.html>; 2011 [retrieved 24.08.10].

[46] Cortes C, Vapnik V. Support-vector networks. In: Saitta L, editor. Machine learning. Boston: Kluwer Academic Publishers; 1995. p. 273–97.

[47] Joachims T. Text categorization with support vector machines: learning with many relevant features, ECML-98. In: 10th European conference on machine learning, 1998. p. 137–42.

[48] McKnight L, Srinivasan P. Categorization of sentence types in medical abstracts, AMIA. Ann Symp Proc 2003:440–4.

[49] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. San Francisco: Morgan Kaufman; 2005.

[50] Artstein R, Poesio M. Inter-coder agreement for computational linguistics. J Comput Linguist 2008;34:555–96.

[51] Di Eugenio B, Glass M. Squibbs and discussions – the kappa statistic: a second look. J Comput Linguist 2004;30:95–101.

[52] Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. J Biomed Inform 2002;35:99–110.

[53] Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12:296–8.

[54] Yates F. Contingency tables involving small numbers and the $\chi 2$ test. Suppl J Roy Stat Soc 1934;1:217–35.

[55] Mowery D, Harkema H, Dowling JN, Lustgarten JL, Chapman WW. Distinguishing historical from current problems in clinical reports – which textual features help? BioNLP 2009:10–8.

[56] Savova G, Chapman W, Zheng J, Crowley R. Anaphoric relations in the clinical narrative: corpus creation. J Am Med Inform Assoc 2011:7.