Semantic Annotation of Clinical Events for Generating a Problem List

Danielle L. Mowery, MS[1], Pamela Jordan, PhD[1], Janyce Wiebe, PhD[1], Henk Harkema, PhD[1], John Dowling, MS MD[1], Wendy W. Chapman, PhD[2]
[1]University of Pittsburgh, Pittsburgh, PA; [2]University of California, San Diego, La Jolla, CA

Abstract

We present a pilot study of an annotation schema representing problems and their attributes, along with their relationship to temporal modifiers. We evaluated the ability for humans to annotate clinical reports using the schema and assessed the contribution of semantic annotations in determining the status of a problem mention as active, inactive, proposed, resolved, negated, or other. Our hypothesis is that the schema captures semantic information useful for generating an accurate problem list. Clinical named entities such as reference events, time points, time durations, aspectual phase, ordering words and their relationships including modifications and ordering relations can be annotated by humans with low to moderate recall. Once identified, most attributes can be annotated with low to moderate agreement. Some attributes – Experiencer, Existence, and Certainty - are more informative than other attributes – Intermittency and Generalized/Conditional - for predicting a problem mention's status. Support vector machine outperformed Naïve Bayes and Decision Tree for predicting a problem's status.

Introduction

In medicine, clinical narratives such as emergency department reports provide a concise overview of the patient's progress with respect to a clinical encounter. The clinical narrative is a flexible medium that supports documentation of signs, symptoms and diseases experienced by a patient accompanied by tests, procedures and treatments administered by care providers to manage the patient's problem status. These narratives have a rich history of use in electronic medical record systems[1] and are written to convey important clinical events that inform clinicians providing quality care. Natural language processing (NLP) is an approach used to identify, encode and extract these clinical events from clinical narratives to support a variety of use cases. NLP techniques can be used to extract patient medication lists[2,3], identify adverse drug effects[4], and generate problem lists[5,6]. Our long-term goal is to develop an NLP system that supports information extraction of clinical named entities and events for patient care environments including automatically generating patient problem lists for care providers and visually displaying medical record information for clinical researchers.

One important step necessary to building an automated NLP system that supports these uses is the development of an annotation schema that explicitly describes the information the NLP system should identify. Typically, humans annotate using the schema and the resulting annotations guide development of an automated extraction system. Before going to the effort of building an NLP system to annotate according to the schema, it is useful to evaluate inter-annotator agreement using the schema and test the informativeness of the schema information for the end goal – if the schema features are not useful for an NLP system, these features should not be encoded. In this paper, we will 1) introduce an annotation schema that supports clinical information extraction, 2) determine how well annotators apply the annotation schema to clinical reports, and 3) evaluate the informativeness of these annotation schema features for predicting a problem mention's status. After revising the schema based on this study, we will annotate a larger corpus and develop NLP methods to extract information according to the schema.

Background

Traditionally, annotation schemas are used to capture information to be manually and/or automatically identified, structured, or extracted from clinical narratives. Researchers have developed these schemas to model semantic information at the document, section, sentence, and mention levels. Mention-level annotations can model a clinical named entity or event at the clause or phrase level, such as the *type of clinical condition* represented by a noun phrase. Researchers develop annotation schemas to model salient clinical named entity and event mentions (NEs) in clinical narratives such as patients, disorders, drugs, procedures, and temporal concepts. In addition to specifying the semantic categories of information to annotate in the report, the schemas often include attributes describing the NEs in context, addressing questions of who, what, when, where, and how. Other schemas aim to encode semantic relationships that occur between mention pairs including is-a and associated-with relationships. In the following section, we will review schemas developed for NEs, attributes, and relationships in clinical text. The works reviewed are not meant to be representative of all annotation schemas, but provide context for the schema we developed, which leveraged existing schemas.

Annotated Clinical Corpora

Several annotated clinical corpora have been developed in recent years to model the information contained in clinical reports, including CLEF[7], i2B2 VA/Challenge[8], and TimeML[9]. As part of a partnership with Royal Marsden Hospital, the CLEF project uses information extraction to support clinical research and evidenced-based medicine[7]. Named entities and events captured include *conditions*, *locus*, *drug*, etc. Condition and Locus have attributes such as *negation* and *laterality*, respectively. Relationships between named entities include coreferring and causal relations. As part of a shared task, the 2010 i2B2 VA/Challenge, discharge summaries were annotated with clinical named entities (*problems*, *tests*, and *treatments*), their assertion attributes (*present*, *absent*, *possible*, etc.), and causal relations (*improves*, *reveals*, etc.)[8]. One of the first efforts to adapt Saurí et al.'s TimeML schema[9] to clinical corpora was undertaken by Savova et al.[10]. Their adaptation includes named entities representing *TIMEX3* and *events*, attributes capturing *tense*, *class*, *degree*, and *modality*, and relationships linking *TLINKS* and *ALINKS.*

Our aim was to develop a schema that integrates named NEs, attributes, and relationships that are important for automatically identifying active problems that should be added to a patient problem list. We borrowed heavily from these existing schemas adding new elements when they did not already exist. We also aimed to align our annotated elements with current annotation initiatives in the NLP community including SHARP's Common Type System[11] and ShARe's Semantic Schema[12] to support the development of a generalizable NLP problem list generator applicable to data from different institutions.

Methods

In the next section, we introduce the annotation schema, describe a pilot study to evaluate the schema, and describe a proof of concept study using attributes in the schema as features for predicting a problem mention's status.

Annotation Schema Introduction

The schema we developed addresses information important for interpreting a patient's clinical conditions: 1) NEs, 2) attributes and their values, and 3) relationships between NEs.

NEs

We developed our annotation schema to encode information related to a patient's disorders; therefore, other NE mentions are only annotated according to the schema if the NE is necessary for interpretation of the disorder.

For each NE, we define boundaries – start and end offsets – for the NE span in the text with square brackets followed by a subscript indicating the annotation type e.g., $[\text{chest pain}]_{CO}$ is a spanned clinical condition mention.

- **Conditions (CO)**: All problems represented by the UMLS semantic group *disorders: signs, symptoms, diagnoses, and test results*. "Patient had minor $[\text{chest pain}]_{CO}$." Condition entities were annotated according to the guidelines described in Chapman et al. 2005[13] and Chapman et al. 2006[14].
- **Reference events (RE)**: *events that place the condition in a particular setting or clinical context.* "Patient $[\text{was referred to cardiology}]_{RE}$ for $[\text{chest pain}]_{CO}$." Reference events are restricted to common care events (admissions, transfers, consults, discharge) and events (motorcycle crash) associated with temporal concepts ("$[\text{CVA}]_{CO}$ from $[\text{motorcycle crash}]_{RE}$ $[\text{in 1990}]_{TI}$") that are useful for determining when the clinical condition occurred.
- **Time points (TP)**: *a particular instance in time.* "$[\text{Three days ago}]_{TP}$ he had a $[\text{stroke}]_{CO.}$"
- **Time durations (TD)**: *an interval or period of time.* "$[\text{For the last three days}]_{TD}$ he denied $[\text{extreme fatigue}]_{CO.}$"
- **Aspectual phase (AP)**: *the stage or phase of the event at a particular point in time* (e.g., *initial, middle, or end).* "The $[\text{onset}]_{AP}$ of her $[\text{nausea}]_{CO}$ occurred after eating a fish dinner." Aspectual phase describes the aspect of the interval representing the life cycle of an NE.
- **Ordering word (OR)**: *an expression positioning two events with respect to each other or a point of perspective in the discourse.* "$[\text{After}]_{OR}$ $[\text{admission}]_{RE}$, she $[\text{vomited}]_{CO}$". A point of perspective can be a reference to other explicit events outside the current sentence boundaries or implicit time perspectives (narrative time, aforementioned time reference) in the discourse.

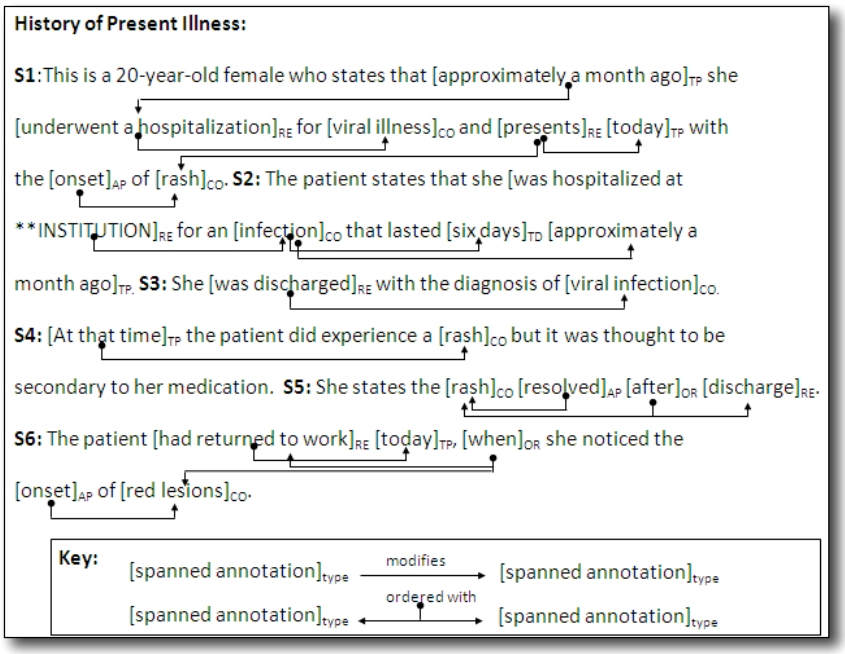

Figure 1. NE mentions spanned with relationships in clinical text.

Figure 1 shows a section of an ED report annotated for these NEs and relationships.

Attributes

Identifying and spanning the NEs mentions in the text alone does not provide the necessary information for understanding the contextual characteristics of the mention in the sentence. In this section, we review the attributes for each NE mention.

Condition Attributes: Every condition mention was annotated with the following attributes and their possible values (bolded values are the values applied to the example sentence):

- Experiencer – who is experiencing the condition.

Ex. The patient's mother had [breast cancer]$_{CO}$. – Experiencer: ***other***, *patient*

- Existence – whether a condition was present or not in the context of the mention.

Ex. He denies [chest pain]$_{CO}$ – Existence: ***no***, *yes*

- Change – whether there is variation in degree or quality of the condition.

Ex. She has had recurrent episodes of [viral meningitis]$_{CO}$ – Change: *unmarked*, *changing*, *unchanging*, *decreasing*, *increasing*, *worsening*, *improving*, ***recurrence***

- Intermittency – whether the condition is episodic in nature.

Ex. White female who complains of [maroon stools]$_{CO}$ two times. – Intermittent: *unmarked*, ***yes***, *no*

- Certainty – the amount of certainty expressed about whether a condition exists or not.

Ex. I have no suspicion for [bacterial infection]$_{CO}$ – Certainty: *unmarked*, ***high***, *moderate*, *low*

- Mental State – whether an outward thought or feeling about a condition is mentioned.

Ex. It seems to me there is some active upper [GI bleeding]$_{CO}$. – Mental State: ***yes,*** *no*

- Generalized/Conditional – whether a condition is in a non-particular or conditional context.

Ex. The patient has [chest pain]$_{CO}$ at rest. – Generalized/Conditional: ***yes,*** *no*

- Relation to Current Visit – position of the condition time interval to the current encounter.

Ex. Past medical history: [Chronic Obstructive Pulmonary Disease]$_{CO}$ – Relation to Current Visit: ***Before***, *Meets_Overlaps*, *After*

For all conditions in which Relation to Current Visit equaled *Before*, we applied the following attributes:

- MagBeforeCurrentVisit – the magnitude of the condition's onset before the current encounter.

Ex. He has had [abdominal pain]$_{CO}$ for the last two days – MagBeforeCurrentVisit: ***2***, *notClear*, *N/A*, *DateGiven*

- UtsBeforeCurrentVisit – the units of the condition's onset before the current encounter.

Ex. He has had [abdominal pain]$_{CO}$ for the last two days – UtsBeforeCurrentVisit: ***days***, *notClear*, *N/A*, *DateGiven*, *hours*, *weeks*, *months*, *years*

Reference Event Attributes: We defined a subset of common key events or reference events describing *where* a condition occurred including common events of ambulatory care visits (admission, discharge, transfer). Our previous study[20] indicated there is also a need for non-clinical events that indirectly link temporal concepts to condition mentions. For instance, in the sentence "[In 2000]$_{TP}$, the patient [had a serious fall]$_{RE}$ resulting in a [shattered knee cap]$_{CO}$.", we know that the knee injury occurred in the year 2000, but the fall (a non-clinical event) is the linguistic link between the condition and temporal concepts. In our annotation schema, reference events do not have attributes, but will eventually be annotated by semantic types e.g., *admission event*.

Point Attributes: Time points provide a reference for *when* a NE occurred. Time points generally refer to the beginning or end of intervals and are sometimes relative to the date of the emergency department visit. The attributes defined for time points will be used to map the time point to the beginning or end of an interval in later processing. We defined the Point attribute values using a subset of the temporal values defined by Zhou et al[15]. and similarly used by Irvine et al[16].

Ex. [Approximately one week ago]$_{TP}$ he had episodes of [fever]$_{CO.}$

- Distance Expressed – whether temporal concept contains a length of time. – ***yes****, no*
- Point Type – what type of temporal concept – *Date and Time,* ***Relative Date and Time****, Fuzzy Time, Point of Perspective, Time Pronoun*

Duration Attributes: Time durations give an indication of *how long* a NE took place. The attributes for durations include common characteristics (beginning, length, and end) used to represent intervals and Duration attribute values are the same as Point attribute values.

Ex. He has had [abdominal pain]$_{CO}$ [for the last two days]$_{TD}$.

- Length Expressed – whether temporal concept contains the length of interval – ***yes****, no*
- Beginning Type – what type of temporal concept is at the start of the interval *–Date and Time,* ***Relative Date and Time****, Fuzzy Time, Point of Perspective, Time Pronoun*
- Ending Type– what type of temporal concept is at the end of the interval *–Date and Time, Relative Date and Time, Fuzzy Time,* ***Point of Perspective****, Time Pronoun*

Aspectual Phase Attributes: Aspectual phase words denote the stage of NEs at a particular time in the narrative. For aspectual mentions, our annotation schema defines attribute values consistent with the TimeML specification[9].

Ex. Her [fever]$_{CO}$ has [abated]$_{AP.}$

- Phase Type –whether beginning, middle or end of event – *Initiation, Continuation,* ***Culmination***

Ordering Word Attributes: Ordering words denote the sequential position of reference events and conditions with respect to one another. We used a simplified set of Allen's temporal intervals[17] similar to Saurí et al.'s TimeML TLinks[9] to annotate the ordering type between NEs. We instructed annotators to determine the ordering type that most closely represents the ordering word (e.g., *follows* is semantically similar to "after") then assign the NEs as arguments 1 and 2 to semantically represent what was meant (e.g., "syncopal episode" after "weak").

Ex. [Syncopal episode]$_{CO}$ [yesterday]$_{TP}$ [after]$_{OR}$ feeling quite [weak]$_{CO}$.

- Ordering Type – temporal position of one entity/event to another – *Precedes, During,* ***Follows***

Relationships

In the early phase of our project, we recognized that the large number of explicit and implicit relationships among NEs could present a cognitive burden on even the most skilled annotator. The focus of our task is to model relationships that describe the condition in a given context relative to a particular place (reference events), time (temporal concepts), stage (aspectual phase), and order (ordering words). As such, we only annotated two relationship types, *modifies* and *orders*, between mentions using four simple and restrictive rules.

- Rule 1: Reference event modifies condition. Only instantiate a condition as a reference event when it serves as a direct link to a clinical condition.

Ex. Patient [presented to the emergency department]$_{RE}$ [chest pain]$_{CO}$ free.

- Rule 2: Temporal concept (Points and Durations) modifies conditions and reference events.

Ex. He has had [abdominal pain]$_{CO}$ [for the last two days]$_{TD}$.

- Rule 3: Aspectual phase modifies conditions and reference events.

Ex. Her [fever]$_{CO}$ has [abated]$_{AP.}$

- Rule 4: Ordering expression orders all combinations of pairs of conditions, reference events and points of perspective with respect to each other.

Ex. [Syncopal episode]$_{CO}$ [yesterday]$_{TP}$ [after]$_{OR}$ feeling quite [weak]$_{CO}$.

The arrows in the following section contain the same meaning as arrows defined in Figure 1 and directional constraints between mentions are illustrated in Figure 2 below.

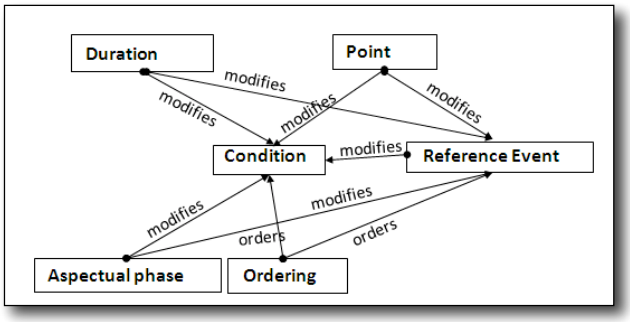

Figure 2. Allowed relationships between NEs.

1. Annotation Study Design

We conducted a pilot annotation study approved by the University of Pittsburgh Institute Review Board. We randomly selected 30 de-identified emergency department (ED) reports from the University of Pittsburgh Medical Center. One author (JD) annotated these reports for clinical conditions. A second annotator reviewed the annotations and came to consensus with JD on any missing or spurious annotations. Five authors (DM, PJ, JW, HH and WC) initially developed the annotation schema based on a literature review of linguistic phenomenon, on existing schemas, and on error analyses of the ConText algorithm[18] and temporal classifiers developed by our lab[19,20]. Once the annotation guidelines were written, two authors (DM and PJ) annotated the remaining NEs, attributes and relationships in the 30 ED reports. Using the Knowtator annotation tool[21], the authors reached consensus using two authors (JW and WC) as adjudicators for any disagreements. The resulting annotations and guidelines were updated iteratively through this process serving as the reference standard and training materials for the annotation study described next.

Medical and nonmedical students were recruited from the University of Pittsburgh using flyers distributed throughout the campus. Over the course of three days, we obtained informed consent, trained subjects about our annotation schema, and reviewed annotation software (Protégé 3.3.1 with Knowtator plugin). In an attempt to reduce the likelihood of annotator fatigue due to the schema's complexity, we assigned the majority attribute value for the previously annotated conditions as default values. Annotators were instructed to change the default value to semantically represent the mention in the text and to annotate additional NEs in the sentence containing the pre-annotated condition as needed. Annotators were given three weeks to independently complete annotation of the 30 ED reports. To determine each annotator's accuracy at the task, we compared each annotator's completed dataset against the reference standard with a python (v 2.5) script we developed, as follows:

NEs

We evaluated the agreement for identifying NEs (other than previously annotated, clinical condition NEs) by assessing annotated mentions against reference standard NE annotations. We considered exact and overlapping span to be true positive NE annotations if the overlapping annotations were assigned the same NE type.

We counted the number of true positives (TP: mention occurs in the reference standard), false positives (FP: mention does not occur in the reference standard) and false negatives (FN: mention was not annotated from the

reference standard). We computed recall to determine the proportion of the reference standard NEs annotators identified and precision to determine the proportion of annotated NEs also generated by the reference standard[22]. We could not count the number of true negative annotations (i.e., a string was correctly not annotated as an NE); therefore, we applied the F_1 score as a surrogate for kappa, since the F_1 score approaches kappa as the number of true negatives become large[22]. We report the mean and standard deviation for each metric.

Attributes

For each leniently matched NE, we applied Cohen's kappa, a chance-corrected agreement measure, between two annotators[23], to attributes with a finite set of values, e.g., *Experiencer* has two values: *patient* or *other*. For NE pair relationships with a lenient match, we measured agreement of their relationship attributes the same way.

Relationships

For relationships, we report the average count of relationships created by annotators. For each relationship that matched the reference standard, we counted the number of true positives (TP: annotated NE pair with arguments occurs in the reference standard), false positives (FP: annotated NE pair with arguments does not occur in the reference standard) and false negatives (FN: annotated NE pair with arguments in the reference standard was not annotated). We calculated recall and precision of relationship pair identification. We also calculated the F_1 score as a surrogate for kappa.

We report the mean and standard deviation for each NE, attribute, and relationship metric.

2. Problem Mention Status Study Design

Using the reference standard generated for attributes of clinical conditions and aspectual phase, we conducted a proof of concept study to evaluate the informativeness of the semantic annotations when predicting a problem mention's status. According to our model of problem lists, a mention of a clinical condition or problem can have one of six possible status labels:

- **Active (A):** a condition mention occurring with high certainty within the patient with an onset within two weeks of the admission and being actively managed during the current episode of care.
- **Inactive (I):** a condition mention chronically experienced by the patient, but not being managed during the current episode of care.
- **Proposed (P):** a condition mention being considered as occurring or diagnosed with less than high certainty.
- **Resolved (R):** a condition mention that occurred during the current episode of care, but was either successfully treated or culminated on its own.
- **Negated (N)**: a condition mention being denied or that never occurred.
- **Other (O):** any other condition mention not classified with the five previous status labels.

Two biomedical informaticians (post doctorates) annotated each condition mention with a status label. One domain expert (physician) adjudicated (Adj) the disagreements, creating the final reference standard. We measured inter-annotator agreement using Cohen's kappa.

We split the dataset into training (70%) and test (30%). Using Weka 3.6.8, we selected three supervised learning classifiers –Decision Tree, Naïve Bayes, and Support Vector Machine– to predict a problem mention's status. We used condition and aspectual phase attributes as input features. We evaluated the condition input features using a feature selection study. Using 10-fold cross validation and the training set, we implemented a best-first, bidirectional search method optimizing accuracy to learn the informativeness of each condition input feature for each classifier. We report the proportion of folds that identified each attribute as informative for high accuracy on the training set. We built a classifier using the full training set and applying only the input features observed as useful in one or more training folds to classify unseen problem mention statuses on the held out test set. We report the performance of the classifier for both training and test sets using Weighted Average Accuracy, Area under the Receiver Operating Curve (ROC), Recall, Precision, and F_1 score.

Results

In this section, we report results of our annotation study and of our problem mention status classification study.

1. Annotation Study

We report annotator agreement with the reference standard for NEs, their attributes, and relationships between them. Of the initial 14 annotators recruited, these results are based on 10 students that completed the annotation study.

NEs

Our dataset is comprised of 30 emergency department reports. The reference standard (RS) has a total of 555 NEs with a distribution of: 283 conditions (51%), 93 reference events (17%), 66 time points (12%), 55 durations (10%), 32 aspectual phase (6%) and 26 ordering words (5%) (Table 1). On average, annotators spanned fewer mentions than the reference standard with mean distribution values of 64 reference events (13%), 59 time points (12%), 37 durations (8%), 31 aspectual phase (6%) and 15 ordering words (3%).

Table 1. Mean proportion of spanned NEs compared to reference standard.

	Reference Events	**Points**	**Durations**	**Aspectual Phase**	**Ordering Words**
RS Counts	93	66	55	32	26
Counts	64.4 +- 23.8	59 +- 13.4	37.4 +- 7.1	31.5 +- 8.3	14.6 +- 9.4
TP	44.6 +- 10.8	47.4 +- 7.3	29.6 +- 5.9	20.9 +- 5.7	8.1 +- 3.8
FP	19.8 +- 16.3	11.6 +-7.8	7.8 +- 2	10.6 +- 4.5	6.5 +- 6.6
FN	48.4 +- 10.8	18.6 +- 7.3	25.4 +- 5.9	11.1 +- 5.7	17.9 +- 3.8
F_1 Score	56.3 +- 9.4	75.9 +- 6.3	63.6 +- 8.0	64.8 +- 13.9	38.2 +- 14.0
Recall	48.0 +- 11.7	73.2 +- 10.5	53.8 +- 10.7	65.3 +- 17.8	31.2 +- 14.7
Precision	72.8 +- 15.4	80.8 +- 9.3	79.1 +- 4.5	66.2 +- 10.8	66.2 +- 22.2

Attributes

Condition Attributes: For condition mentions, the average kappa agreement between an annotator and the reference standard varied from low kappa for *Intermittency*: 0.39 +- 0.1 and *Generalized or Conditional*: 0.46 +- 0.3 to moderate kappa for *Magnitude before Current Visit*: 0.5 +- 0.2, *Certainty*: 0.52 +- 0.1, *Units before Current Visit*: 0.56 +- 0.2, *Mental State*: 0.59 +- 0.2, *Change*: 0.63 +- 0.1, *Relation to Current Visit*: 0.64 +- 0.1 to high kappa for *Existence*: 0.8+-0.1 and *Experiencer*: 1.0 +- 0.

Point Attributes: For point mentions, annotators correctly identified an average of 47 matches with the reference standard. Annotators achieved moderate kappa for *Type*: 0.6 +- 0.2 and *Distance Expressed*: 0.7 +- 0.2 attribute values.

Duration Attributes: For duration mentions, annotators correctly identified an average of 30 matches with the reference standard. Annotators achieved low to moderate kappa for *Beginning Type*: 0.40 +- 0.2, *Ending Type*: 0.50 +- 0.3, and *Length Expressed*: 0.70 +- 0.2.

Aspectual Phase Attributes: For aspectual phase mentions, annotators correctly an average of 21 matches with the reference standard. Annotators achieved high kappa for *Phase Type*: 0.96 +- 0.3.

Ordering Word Attributes: For ordering word mentions, annotators correctly identified an average of 8 TP matches with the reference standard. Annotators averaged moderate agreement for *Ordering Type:* 0.4 +- 0.4.

Relationships

Reference Event Relationships: On average, annotators correctly identified 45 (59%) modifying relationships. For correctly identified modifying relationships, annotators achieved high recall and precision identifying the events condition mentions being modified (Table 2).

Point Relationships: On average, annotators correctly identified 47 (72%) modifying relationships. For correctly identified modifying relationships, annotators produced substantial recall and precision for identifying reference events and conditions being modified.

Duration Relationships: On average, annotators correctly identified 30 (53%) modifying relationships. For correctly identified modifying relationships, annotators produced substantial recall and precision for identifying reference events and conditions being modified.

Aspectual Phase Relationships: On average, annotators correctly identified 21 (60%) modifying relationships. For correctly identified modifying relationships, annotators produced high recall and precision for identifying reference events and conditions being modified.

Ordering Word Relationship: On average, annotators correctly identified 8 (27%) ordering relationships. For correctly identified ordering relationships, annotators had low recall and precision identifying events being ordered.

Table 2. Mean relationships and arguments identified by annotators.

NEs	Reference Events	Points	Durations	Aspectual Phase	Ordering Words
RS	76	65	57	35	30
Counts	44.6 +- 10.8	47.4 +- 7.3	29.6 +- 5.9	20.9 +- 5.7	8.1 +- 3.8
F_1 Score	94.6 +- 4.9	77.1 +- 4.7	88.1 +- 5.4	91.5 +- 2.9	19.5 +- 9.4
Recall	92.1 +- 8	77.6 +- 6.8	88.6 +- 4.9	91.9 +- 4.2	16.0 +- 9.4
Precision	97.6 +- 1.8	76.9 +- 4.1	87.6 +- 6.5	91.3 +- 3.6	35.3 +- 26.9

2. Problem Mention Status Study

Kappa agreement between pairs was A1-A2 (23.6%), A1-Adj (33.4%), and A2-Adj (77.3%). The most prevalent status was Active among annotators (Table 3). The majority of disagreements between A1-A2 were Inactive/Active.

Table 3. Count and prevalence (%) of status label by annotator.

	Active (A)	Inactive (I)	Proposed (P)	Resolved (R)	Negated (N)	Other (O)
A1	110 (39%)	101 (36%)	5 (2%)	29 (2%)	31 (11%)	7 (2%)
A2	198 (70%)	28 (10%)	7 (2%)	0 (0%)	28 (10%)	22 (8%)
Adj	181 (64%)	21 (7%)	7 (2%)	22 (8%)	26 (9%)	25 (9%)

The most prevalent attribute values observed for conditions in the reference standard were *Experiencer*: patient (98%), *Existence*: yes (89%), *Certainty*: unmarked (95%), *Mental State*: no (95%), *Intermittency*: unmarked (83%), *Change*: unmarked (82%), *Generalized/Conditional*: no (88%), *Relation to Current Visit*: Meets_Overlaps (63%), *Magnitude Before Current Visit* (not shown): notClear (55%) and *Units Before the Current Visit* (not shown): notClear (55%) (Table 4). The attribute values reflect the prevalence of feature attribute values for classification.

Table 4. Reference standard - distribution of condition attribute values used in feature vectors.

Experiencer	Existence	Certainty	Mental State	Intermittency	Change	Generalized/ Conditional	Relation to Current Visit
yes (98%)	*yes* (89%)	*unmarked* (95%)	*yes* (95%)	*unmarked* (83%)	*unmarked* (82%)	*yes* (88%)	*Meet/Overlap* (63%)
no (2%)	*no* (11%)	*moderate* (3%)	*no* (5%)	*yes* (16%)	*worsening* (5%)	*no* (12%)	*Before* (29%)
		low (>1%)		*no* (>1%)	*unchanging* (4%)		*After* (8%)
		high (<1%)			*improving* (3%)		
					increasing (3%)		
					decreasing (2%)		
					recurrence (1%)		

We observed condition attributes, *Experiencer* and *Existence,* are consistently 100% informative for asserting a problem mention's status among classifiers (Table 5). Naïve Bayes and Support Vector Machine determined all attributes relevant for at least 1 fold. In contrast Decision Tree, only determined 5 of 10 attributes relevant.

Table 5. Count (#) of Folds/10 that an attribute was determined relevant.

Condition attributes	Decision Tree	Naïve Bayes	Support Vector Machine
Experiencer	10 (100%)	10 (100%)	10 (100%)
Existence	10 (100%)	10 (100%)	10 (100%)
Change	0 (0%)	8 (80 %)	10 (100%)
Intermittency	0 (0%)	3 (30%)	4 (40%)
Certainty	7 (70%)	8 (80%)	10 (100%)
Mental State	0 (0%)	2 (20%)	9 (90%)
Generalized/Conditional	0 (0%)	1 (10%)	3 (30%)
Relation to Current Visit	10 (100%)	4 (40%)	9 (90%)
Magnitude&Units > 2 wks	0 (0%)	6 (60%)	6 (60%)
Aspectual Phase	1 (10%)	8 (80%)	10 (100%)

Our training set of 198 (70%) conditions had a distribution of Active 127 (64%), Inactive 15 (8%), Proposed 6 (3%), Resolved 15 (8%), Negated 18 (9%), and Other 17 (9%); our test set of 85 (30%) conditions had a distribution of Active 54 (64%), Inactive 6 (7%), Proposed 2 (3%), Resolved 7 (8%), Negated 8 (9%), and Other 8 (9%). All classifiers outperformed a majority class baseline (Active: 64% Overall Accuracy) in Table 6. For Weighted Average Accuracy, the test set was between 4-9 points lower than the training set among classifiers. For Weighted Accuracy and F_1 Score, Support Vector Machines demonstrated higher performance over Decision Tree and Naïve Bayes. Performances were higher for *Active* and *Negated* and lower for *Inactive* and *Resolved* among classifiers.

Table 6. Classifier performances on training and test data.

Classifier	**Status**	**Wt. Accuracy**		**ROC**		**Recall**		**Precision**		**F_1 Score**	
		Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Decision Tree	A			80.9	83.8	93.7	90.7	77.8	77.8	85.0	83.8
	I			79.3	89.9	0.0	0.0	0.0	0.0	0.0	0.0
	P			82.0	67.8	33.3	50.0	50.0	50.0	40.0	50.0
	R			56.9	73.4	0.0	0.0	0.0	0.0	0.0	0.0
	N			99.0	98.1	100.0	100.0	0.9	72.7	94.7	84.2
	O			94.2	75.4	94.1	62.5	80.0	55.6	86.5	58.8
Wt. Ave		78.3	74.1	81.8	83.5	78.3	74.1	66.5	62.7	71.8	67.9
Naïve Bayes	A			85.6	84.1	89.0	83.3	77.9	75.0	83.1	78.9
	I			83.4	87.6	6.7	0.0	9.1	0.0	7.7	0.0
	P			95.2	98.8	16.7	0.0	50.0	0.0	25.0	0.0
	R			85.1	80.5	13.3	0.0	40.0	0.0	20.0	0.0
	N			99.4	100.0	94.4	100.0	89.5	72.7	91.9	84.2
	O			97.8	86.4	70.6	50.0	75.0	50.0	72.7	50.0
Wt. Ave		73.7	67.1	88.0	86.1	73.7	67.1	69.8	59.2	70.7	62.8
Support Vector Machine	A			81.7	76.0	92.9	87.0	84.9	81.0	88.7	83.9
	I			85.6	81.0	13.3	0.0	100.0	0.0	85.6	0.0
	P			99.3	99.1	83.3	50.0	71.4	50.0	76.9	50.0
	R			87.9	76.8	46.7	28.6	70.0	50.0	56.0	36.4
	N			99.7	99.1	100.0	100.0	90.0	72.7	94.7	84.2
	O			99.2	78.0	100.0	75.0	85.0	60.0	91.9	66.7
Wt. Ave		84.3	75.3	86.1	79.3	84.3	75.3	85.0	69.3	81.8	71.7

Discussion and Future Work

1. Annotation Study: We introduced an annotation schema for clinical information extraction of events for generating an accurate problem mention status. We learned that agreement for annotation condition attributes ranges from moderate to high. Annotators had some difficulty identifying other clinical named entities and their relationships suffering from low to moderate recall; annotators had moderate agreement for other clinical named entities attributes. This observation is not surprising as the literature shows agreement suffers beyond 2 categories especially for less prevalent categories[24]. Indeed, a study of the CLEF schema reports moderate F_1 scores in the 60s for entity and relationship annotations from clinical narratives[7]. We plan to increase training, apply NLP system pre-annotations, and use eHOST's Oracle function to improve recall and classification of NEs in future studies[25].

2. Problem Mention Status Study: From our feature selection study, we learned that attributes like *Experiencer*, *Existence*, and *Certainty* are consistently more informative than other attributes for predicting mention status among classifiers. From our classification study, we observed that classifiers (Naïve Bayes and Support Vector Machine) that use rare occurring attributes like *Change, Mental State*, and *Intermittency* perform better than a classifier (Decision Tree) that does not use them. We suspect our classifiers performed poorly predicting status labels for *Inactive*, *Resolved*, and *Proposed* due to subtle differences in definition between status labels (Inactive and Resolved) and few instances in the dataset. In terms of comparable studies, like the i2B2 assertion classification, other researchers have demonstrated adding lexical, syntactic, section, and other semantic annotations can boost performance[26]. We plan to add such annotations including Unified Medical Language System concept unique identifiers and discourse annotations to our schema before applying our experiments to the SHARP and ShARe corpora. We will expand our study to other report types such as discharge summaries, radiology, electrocardiograms, echocardiograms, and progress notes. We are actively annotating document-level problem annotations and their statuses for these corpora.

Acknowledgements

We thank the NLM and NIH for funding this study with grants 5T15LM007059 and R01LM009427, our anonymous reviewers and annotators including, but not limited to Mike Conway, Son Doan, and Mindy Ross.

References

1. Tange HJ, Hasman A, de Vries Robbé PF, Schouten HC. Medical narratives in electronic medical records. Int J Med Inform. 1997;46(1):7-29.
2. Hua X, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc. 2009;17(1):19-24.
3. Sohn S, Murphy S, Masanz J, Kocher J, Savova G. Classification of Medication Status Change in Clinical Narratives. AMIA Annu Symp Proc. 2010:762-7.
4. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al. Extraction of Adverse Drug Effects from Clinical Records: IOS Press: IMIA and SAHIA; 2010.
5. Meystre S, Haug P. Automation of a Problem List using Natural Language Processing. BMC Medical Informatics and Decision Making 2005; **5**(30).
6. Solti I, Aaronson B, Fletcher G, Solti M, Gennari J, Cooper M, Payne, T. Building an Automated Problem List based on Natural Language Processing: Lessons Learned in the Early Phase of Development. 2008: 687–691.
7. Roberts A, Gaizauskas R, Hepple M, Davis N, Demetriou G, Guo Y, et al. The CLEF Corpus: Semantic Annotation of Clinical Text. AMIA Annu Symp Proc. 2007:625-9.
8. Uzuner O, South B, Shen S, DuVall S. 2010 i2b2/VA Challenge on Concepts, Assertations, and Relations in Clinical Text. J Am Med Inform Assoc. 2011;18:552-6.
9. Saurí R, Littman J, Knippen B, Gaizauskas R, Setzer A, Pustejovsky J. TimeML Annotation Guidelines Version 1.2.1.2006: Available from: http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf.
10. Savova G, Bethard S, Styler W, Martin J, Palmer M, Masanz J, Ward W. Towards Temporal Relation Discovery from the Clinical Narrative. AMIA Annu Symp Proc. 2009: 568-572.
11. Wu ST, Kaggal VC, Dligach D, Masanz JJ, Chen P, Becker L, Chapman WW, Savova GK, Liu H, Chute CG. A common type system for clinical natural language processing. J Biomed Semantics. 2013;4(1).
12. Elhadad N, Chapman WW, O'Gorman T, Palmer M, Savova G. The ShARe Schema for the Syntactic and Semantic Annotation of Clinical Texts. Under Review.
13. 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.
14. Chapman W, Dowling J. Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. J Biomed Inform. 2006;39(2):196-208.
15. Zhou L, Melton GB, Parsons S, Hripcsak G. A temporal constraint structure for extracting temporal information from clinical narrative. J Biomed Inform. 2006;39(4):424-39.
16. Irvine A, Haas S, Sullivan T. TN-TIES: A System for Extracting Temporal Information from Emergency Department Triage Notes. AMIA Annu Symp Proc. 2008:328–32.
17. Allen J. Towards a General Theory of Action and Time. Artif Intell Med. 1984;23(2):123-54.
18. Chapman WW, Chu D, Dowling JN. ConText: An Algorithm for Identifying Contextual Features from Clinical Text. Association for Computational Linguistics; 2007; Prague, Czech Republic.
19. Mowery D, Harkema H, Chapman WW. Temporal Annotation of Clinical Text. BioNLP Workshop 2008: Current Trends in Biomedical Natural Language Processing; Columbus, OH. 2008;1-2.
20. Mowery D, Harkema H, Dowling JN, Lustgarten JL, Chapman WW. Distinguishing Historical from Current Problems in Clinical Reports -- Which Textual Features Help? BioNLP. 2009;10-18.
21. Ogren P. Knowtator: a protege plug-in for annotated corpus construction. Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology; New York, New York: Association for Computational Linguistics; 2006:273-5.
22. Hripcsak G, Rothschild AS. Agreement, the F-Measure, and Reliability in Information Retrieval. J Am Med Inform Assoc. 2005;12:296-8.
23. Artstein R, Poesio M. Inter-coder Agreement for Computational Linguistics. Comp Ling. 2008;34(4):555-96.
24. Poesio M, Vieira, R. A corpus-based investigation of definite description use. Comp Ling. 1998;24(2):183-216.
25. South BR, Shuying S, Leng J, Forbush, TB, DuVall SL, Chapman WW. A prototype tool set to support machine-assisted annotation. BioNLP. 2012;130-139.
26. Clark C, Aberdeen J, Coarr, Tresner-Kirsch D, Wellner B, Yeh A, Hirschman L. MITRE system for clinical assertion status classification. J Am Med Inform Assoc. 2011;18:563-567.