Journal of Biomedical Informatics 42 (2009) 839–851

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports

Henk Harkema [a,*], John N. Dowling [a], Tyler Thornblade [b], Wendy W. Chapman [a]

[a] Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15260, USA
[b] Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260, USA

ARTICLE INFO

Article history:
Received 9 June 2008
Available online 10 May 2009

Keywords:
Natural language processing
Negation
Temporality
Clinical reporting

ABSTRACT

In this paper we describe an algorithm called ConText for determining whether clinical conditions mentioned in clinical reports are negated, hypothetical, historical, or experienced by someone other than the patient. The algorithm infers the status of a condition with regard to these properties from simple lexical clues occurring in the context of the condition. The discussion and evaluation of the algorithm presented in this paper address the questions of whether a simple surface-based approach which has been shown to work well for negation can be successfully transferred to other contextual properties of clinical conditions, and to what extent this approach is portable among different clinical report types. In our study we find that ConText obtains reasonable to good performance for negated, historical, and hypothetical conditions across all report types that contain such conditions. Conditions experienced by someone other than the patient are very rarely found in our report set. A comprehensive solution to the problem of determining whether a clinical condition is historical or recent requires knowledge above and beyond the surface clues picked up by ConText.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

In this paper we introduce and evaluate an algorithm called ConText for determining whether a clinical condition is negated, hypothetical, historical, or experienced by someone other than the patient. This type of algorithm has potential to substantially improve precision for information retrieval and extraction from clinical records. For instance, a query for patients with a diagnosis of pneumonia may return false positive records for which pneumonia is mentioned but is negated (e.g., "ruled out pneumonia"), experienced by a family member (e.g., "family history of pneumonia"), or occurred in the past ("past history of pneumonia").

ConText is a simple algorithm that can be easily integrated in applications that index clinical conditions. It is derived from the NegEx algorithm for identifying negated findings and diseases in discharge summaries [1]. NegEx uses regular expressions to identify the scope of trigger terms that are indicative of negation such as "no" and "ruled out." Any clinical conditions within the scope of a trigger term are marked as negated.

The content and evaluation of the ConText algorithm in this paper extends previous work on NegEx in various important and practically relevant ways. First, although ConText borrows the approach based on trigger terms and regular expressions from NegEx, it employs a different definition for the scope of trigger terms. Second, as illustrated above, clinical conditions can be modified by several contextual properties that are relevant for clinical NLP applications; ConText identifies three contextual values in addition to NegEx's negation: hypothetical, historical, and experiencer. In this paper we will address the question of whether the simple approach based on regular expressions that works well for capturing negation in clinical reports can be successfully transferred to other contextual properties. Third, the electronic medical record holds clinical documents of various types. The difference in content and style among these types of report may be considerable. In this paper we evaluate ConText on a clinical document set comprised of six report types. Therefore, the evaluation results will also show to what extent the regular expression approach to identifying contextual properties is portable among report types. As a final contribution, the paper provides an overview of the prevalence of the four contextual properties in each of the six report types. These prevalence numbers can be taken as an indication of how useful contextual property identification is for each of the report types.

2. Background

Clinical documents are a valuable source of information for detection and characterization of outbreaks, decision support, recruiting patients for clinical trials, and translational research,

* Corresponding author. Address: Department of Biomedical Informatics, University of Pittsburgh, 200 Meyran Avenue, VALE M-183, Pittsburgh, PA 15260, USA. Fax: +1 412 647 7190.
E-mail address: heh23@pitt.edu (H. Harkema).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.05.002

because they contain information regarding signs, symptoms, treatments, and outcomes. For example, radiology, surgical pathology, molecular pathology, cytogenetic, and flow cytometry reports contain valuable information for translational cancer research that can be used for epidemiologic and descriptive studies and discovery of new relationships that impact diagnosis and prognosis or treatment. Most of the information contained in clinical documents is locked in free-text format and must be encoded in a structured form to be useful for these applications.

The biomedical informatics community has produced decades of research resulting in dozens of applications for indexing, extracting, and encoding clinical conditions from clinical documents stored in the electronic medical record [2]. Most applications have focused on identifying individual conditions at the sentence level (e.g., identifying the condition Dyspnea in the sentence "Patient complains of shortness of breath."), and a few systems attempt to represent a fairly complete semantic model of the conditions. For example, MedLEE [3], MPLUS [4], MEDSYNDIKATE [5], the Multi-threaded Clinical Vocabulary Server (MCVS) [6], and a radiology report encoding system developed by Taira and colleagues [7] all identify not only the condition but also modifying information such as anatomic location, negation, change over time, and severity.

Most medical language processing applications index or extract individual clinical conditions but do not model much information found in the context of the condition. For instance, MetaMap [8], available from the National Library of Medicine and used in several clinical applications, indexes UMLS concepts in text, and has been used to index symptoms, signs, and diagnoses described in clinical reports [9,10]. Conditions indexed by MetaMap are largely comprised of contiguous text contained in simple noun phrases. Other research groups have developed similar systems that handle inflectional and derivational variants, synonymy, and even polysemy [11]. To be useful for clinical applications such as looking for genotype/phenotype correlations, retrieving patients eligible for a clinical trial, or identifying disease outbreaks, simply identifying clinical conditions in the text is not sufficient—information described in the context of the clinical condition is critical for understanding the patient's state.

Others have developed applications for modifying biomedical text with contextual information. Light and colleagues [12] determine whether concepts are described as facts or as speculation based on the context in which the concept occurs in biomedical articles. Mizuta and colleagues [13] developed a classifier for rhetorical zones in biology articles to provide useful context for information extraction. Medlock and Briscoe [14] applied a weakly supervised learning algorithm for classifying hedges in the biomedical literature. Medical language processing systems that encode clinical information from textual reports [3–5,15] extract not only the clinical condition but also contextual properties such as certainty, anatomic location, change over time, and severity that modify the condition.

Negation is probably the most important contextual feature in clinical reports. One study showed that approximately half of the conditions indexed in dictated reports are negated [16]. Another study showed that negation status was the most important feature for classifying patients based on whether they had an acute lower respiratory syndrome; including negation status contributed significantly to classification accuracy [17]. Several methods exist for determining whether a condition is negated [1,18–20]. To our knowledge, NegEx [1] is the only stand-alone negation detection system that is freely available for use by others (for several years a Python version has been available by emailing the authors. Now several Python implementations and a Java version can be downloaded from http://code.google.com/p/negex). NegEx has been used by a variety of biomedical indexing applications [21–25], indicating the need for a stand-alone processing component that can be easily deployed by others. NegEx is also being integrated with the NLM's MetaMap indexing system and will be distributed in MetaMap's next release. ConText [26] is an extension of NegEx that also addresses other contextual properties, currently including whether a condition is historical, hypothetical, or experienced by someone other than the patient. Like NegEx, ConText can be integrated with any application that indexes clinical conditions from text, because it does not rely on any level of syntactic or semantic analysis.

Various adaptations and reimplementations of NegEx as well as several other algorithms for negation detection have been integrated in NLP applications that process a variety of different clinical report types (e.g., mammography reports [27], pathology reports [23], clinical practice guidelines [28], and discharge summaries [1]). However, comparative studies that examine directly NegEx's scope and evaluate a version of a contextual feature detection algorithm on a set of different report types to assess its generality, as in the present paper, have been lacking. We are aware of one other study in this vein: Uzuner et al. [29] specify a rule-based and a statistical method for detecting negation and experiencer status of conditions in radiology reports and find that these methods can be directly applied to discharge summaries with similar performance results.

3. Methods

The ConText algorithm uses regular expressions over pre-indexed clinical conditions and specific sets of words in text to identify conditions that are negated, hypothetical, historical, or experienced by someone other than the patient. Previous NegEx results show that a regular expression-based approach works well for detecting negation in discharge summaries [1]. The objective of this paper is to assess the versatility of this approach along two dimensions: how well does it work for contextual properties other than negation and how well does it work across diverse report types? To answer these questions we ran ConText over a corpus containing six types of reports commonly stored in the electronic medical record and evaluated its performance for the contextual properties addressed by the algorithm.

In this section we first describe the ConText algorithm. We also discuss the other methodological aspects of our study, including the dataset, reference standard, and evaluation methods.

3.1. ConText algorithm

The idea underlying NegEx is that a clinical condition in text is affirmed by default and that a departure from the default value, i.e., the condition is absent, can be inferred from simple lexical clues occurring in the context of the condition. ConText takes this idea and extends it to other contextual properties.

ConText is a regular-expression based algorithm that searches for trigger terms preceding or following the indexed clinical conditions. If a condition falls within the scope of the trigger term, ConText changes the default value to the value indicated by that trigger term. Fig. 1 illustrates ConText's actions on the sentence "No history of chest tightness but family history of CHF."

3.1.1. Contextual properties

ConText determines the values for three contextual properties of a clinical condition: Negation, Temporality, and Experiencer. The contextual property negation specifies the status of the clinical existence of a condition. The default value of this property is *affirmed*. If a clinical condition occurs within the scope of a trigger term for negation, ConText will change the default value to *negated*. For example, in the sentence "The patient denies any nausea," the value of negation for the condition "nausea" will be *negated*.

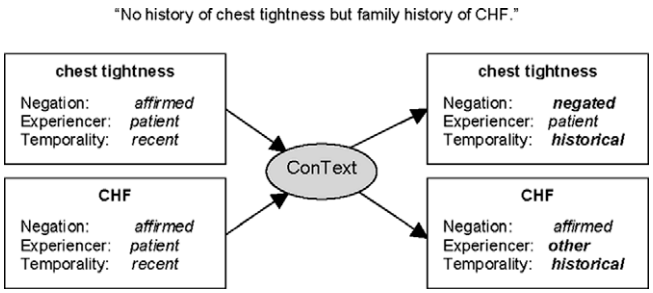

Fig. 1. ConText changes the default values for contextual features based on words in the sentence. Bold values on the right were changed by ConText.

The contextual property temporality places a condition along a simple time line. The default value of temporality is *recent*. Given appropriate trigger terms, ConText can change the value of this property to either *historical* or *hypothetical*. In our current annotation schema the value *historical* is defined to apply to conditions beginning greater than 2 weeks previous to the visit or procedure that is documented in the clinical note. The value *hypothetical* covers all conditions that temporally are neither recent nor historical. A typical example of a hypothetical condition would be "fever" in a sentence such as "Patient should return if she develops fever."

Finally, the contextual property experiencer describes whether the patient or someone else experiences the condition. The default value is *patient*, which, in the presence of a trigger term, can be changed to *other*. For example, in the sentence "The patient's father has a history of CHF," the value of experiencer for the condition CHF is *other*. The particular choice of contextual properties and their values in ConText is based on their usefulness for disease classification in a biosurveillance application [17].

3.1.2. Trigger terms, pseudo-trigger terms, and termination terms

As indicated above, trigger terms prompt ConText to change the default value of a contextual property for a condition, provided the condition falls within the scope of the trigger term. For each non-default value of a contextual property ConText maintains a separate list of trigger terms, i.e., the values *negated*, *historical*, *hypothetical*, and *other* are all triggered by their own set of trigger terms. For instance, the set of trigger terms for *negated* includes terms like "no" and "denies," for *hypothetical*, "if" and "should," for *historical*, "history" and "status post," and for *other*, "family history" and "mother's." The total number of trigger terms used by the current version of ConText is: 143 for *negated*, 10 for *historical*, 11 for *hypothetical*, and 26 for *other*. ConText also implements a few regular expressions to capture explicit temporal expressions such as "1-week history" or "three months ago." If the temporal value contained in the regular expression is greater than 14 days, the condition is classified as *historical*.

ConText also uses pseudo-trigger terms for terms that contain trigger terms but do not act as contextual property triggers. For example, the temporality trigger "history" often denotes a temporality value of *historical* but also appears in text without affecting the temporality value, as in "I performed the patient's physical and history exam." To avoid false positives, "History exam" is included in the list of pseudo-triggers for *historical*. In the current version of ConText there are 17 pseudo-triggers for *negated* (e.g., "no increase," "not cause"), 17 pseudo-triggers for *historical* (e.g., "social history," "poor history"), four pseudo-triggers for *hypothetical* (e.g., "if negative," "know if"), and 18 pseudo-triggers for *other* (e.g., "by her husband," "by his brother").

The default scope of a trigger term includes all clinical conditions following the trigger term until the end of the sentence, but this scope can be overridden. Certain words, or termination terms, in a sentence can signal the end of the scope of a trigger term. For example, in the sentence "History of COPD, presenting with shortness of breath," the trigger term "history" makes the condition COPD historical, but the term "presenting" indicates that the physician has switched to describing the current patient visit. Therefore, ConText treats the term "presenting" as a termination term ending the scope of the trigger term "history," and the condition "shortness of breath" will be classified as *recent* rather than *historical*. Termination terms have been assembled into 12 conceptual groups: Presentation, Patient, Because, Diagnosis, ED, Etiology, Recent, Remain, Consistent, Which, And, and But terms.

Although trigger terms are unique to the contextual feature being identified, termination terms may be common to multiple contextual properties. For example, Presentation terms, which include "presenting" and "presents," are termination terms for both *historical* (see the example above) and *hypothetical*, as illustrated in the sentence "Mother has CHF and patient presents with chest pain." Experiencer for CHF should be *other*, but experiencer for Chest Pain should be *patient*. Table 1 provides more details about the termination terms included in the current version of the ConText system.

The initial lists of trigger terms, pseudo-trigger terms, and termination terms for ConText were gathered manually from a development set of emergency department reports.

3.1.3. Scope of trigger terms

A trigger term affects the values of the contextual properties of all conditions in its scope. In the general case, the scope of a trigger term extends to the right of the trigger term and ends at a termination term or at the end of the sentence, whichever comes first. Besides this standard definition, ConText contains three alternative definitions of scope that apply to specific sets of trigger terms. First, for the value *historical*, there is a small set of trigger terms whose scope is restricted to one condition to the right of the trigger term. These trigger terms are "pre-existing," "status post," and "s/p." Second, for the value *negated*, there is a set of 14 "left-looking" trigger terms or post-triggers. The scope of these trigger terms runs from the trigger term leftward to the beginning of the sentence, and can be terminated by any regular, intervening termination term. The list of post-triggers includes terms such as "is ruled out," "are not seen," and "negative." Third, if a trigger term for *historical* occurs within the title of a section, its scope extends throughout the entire section. Operation of this rule requires that sections and section titles have been identified and demarcated in text. This scope rule applies, for example, to mark all conditions in a "Past Medical History" section as *historical*.

Table 1
ConText's termination terms. The first column in the table below lists the classes of termination terms used by ConText, including the number of terms in each class. The second column gives examples of termination terms in each class. The last four columns of the table indicate which contextual property values (*negated*, *historical*, *hypothetical*, and *other*) rely on which class of termination terms.

Term class	Example terms	Applies to			
		neg.	*hist.*	*hyp.*	*other*
Presentation (15)	Presents, comes in with	√	√		√
Patient (5)	Patient, his			√	√
Because (2)	Since, because	√		√	
Diagnosis (1)	Diagnosis			√	
ED (2)	Emergency department, ED		√		
Etiology (23)	Origin of, secondary to	√			
Recent (1)	Recent		√		
Remain (2)	Remains, remained	√			
Consistent (1)	Consistent with		√		
Which (1)	Which				√
And (2)	And, so	√			
But (8)	But, aside from	√			

3.1.4. Algorithm

The input to the ConText algorithm is a sentence (or a text with marked-up sentence boundaries) in which clinical conditions have been indexed and default values have been assigned to their contextual properties. The output of ConText is the input sentence in which the values of the contextual properties of the conditions have been updated according to the following simple algorithm:

1. Mark up all trigger terms, pseudo-trigger terms, and termination terms in the sentence.
2. Iterate through the trigger terms in the sentence from left to right:
   a. If the trigger term is a pseudo-trigger term, skip to the next trigger term.
   b. Otherwise, determine the scope of the trigger term and assign the appropriate contextual property value to all indexed clinical conditions within the scope of the trigger term.

Fig. 2 illustrates how ConText uses trigger and termination terms to determine the values for contextual properties in the sentence "Past history of pneumonia presenting today with cough and fever."

The implementation of the algorithm uses a set of regular expressions over trigger terms, termination terms, and "end-of-sentence" annotations which match the scope of trigger terms. The algorithm steps through the conditions inside the scope of a trigger term and changes the value of the contextual properties of the conditions as indicated by the type of trigger term. We have implemented ConText as a processing resource within the GATE NLP framework, using GATE'S JAPE grammar formalism to specify the regular expressions and associated actions [30]. The trigger terms etc. in text are marked up through lexical look-up. In order to deal with term variation, the term lists contain all relevant variants of a term. For example, the Presentation termination term list includes the terms "presented," "presents," and "presenting." The description of the ConText algorithm, the lists of terms, the GATE processing resource, and a Python implementation are available from the NegEx Google Code page, which is located at http://code.google.com/p/negex/.

In an initial study reported in [26], we showed that ConText performed on emergency department reports with high recall and precision for negation (97%, 97%), moderate recall and precision for temporality (hypothetical) (83%, 94%), and fair recall and precision for temporality (historical) (67%, 74%) and experiencer (50%, 100%). In the present paper we expand our evaluation of the efficacy of employing surface trigger terms to identify values of contextual properties and we explore ConText's portability between clinical reports of different types.

3.1.5. ConText and NegEx

ConText shares with NegEx the idea that contextual properties can be assigned a default value, and that departures from the default value are indicated by specific words in the text. ConText differs from NegEx in the way the scope of a trigger term is defined. In the NegEx algorithm the scope of a trigger term is a priori restricted to a window of six tokens (where multiword concepts count as one token) following the trigger term (preceding six tokens for post-triggers). If any of these six tokens is a termination term or if the window includes the end of the sentence, the scope ends at that point. As described above, ConText has a more liberal definition of scope. There is no six-token window: the scope ends with a termination term or at the end of the sentence, however far removed from the trigger term. Experimental results show that extending the scope in this way improves performance. This will be discussed in more detail in Section 4.

Fig. 2. ConText changes the default value of a contextual property for all clinical conditions (in bold) within the scope of the trigger term, which is usually until the end of the sentence or at a termination term. In this example, the trigger term "history" indicates that the clinical conditions up until the termination term "presenting" are *historical* rather than the default *recent*.

3.2. Datasets of six report types

We applied ConText to clinical conditions indexed within six types of clinical reports stored in our Electronic Medical Record: radiology, emergency department, surgical pathology, echocardiogram, operative procedures including GI endoscopy and colonoscopy, and discharge summaries. The report types we selected all describe clinical conditions experienced or manifested by a patient, including laboratory, radiology, and physical findings, symptoms, and diagnoses, making them potential candidates for ConText's processing. However, the report types we selected differ in purpose and in style, constituting in reality very different genres. Some qualitative differences between report types include a focus on actions (such as operative procedures) or descriptions, the number of numeric measurements described, the extent to which physicians generating the reports attempt to proffer a diagnosis or just describe findings, and the proportion of conditions the patient does not have that are listed in the report (ruling out worrisome conditions). We measured some surface differences among the report types, including the length of the reports in words, the length of the sentences, and the number of clinical conditions they contain.

With IRB approval, we collected a randomized set of 240 deidentified reports stored in the University of Pittsburgh Medical Center's MARS repository between March and April, 2007. The dataset contained 40 of each of the six report types. We used a subset of 120 reports (20 of each report type) as a development set to assess ConText's performance and make changes based on an error analysis. We used the second set of 120 reports as an independent test set to evaluate the revised version of ConText.

3.3. Reference standard annotations

A physician board-certified in internal medicine and infectious diseases with 30 years of clinical experience and extensive annotation practice (author JND) provided reference standard annotations. He annotated all clinical conditions in the data set, applying an annotation schema we previously developed and evaluated [31]. As described in the annotation guidelines, clinical conditions included signs, symptoms, and diseases but did not include demographics or risk factors. He annotated findings with qualitative values (e.g., "low blood pressure") but did not annotate findings with quantitative values (e.g., "blood pressure 90/55"). For every annotated condition, he assigned values to the three contextual features. Because a single annotator is prone to error and annotation fatigue, we ran ConText over the physician's annotated conditions in the development set and presented him with the contextual property values on which he and ConText disagreed. For each disagreement the physician did not know which contextual feature value was assigned by him and which was assigned by ConText, and he selected the value that he believed was correct. We used his corrected assignments as the reference standard for the development set.

For the test set, we used a modified form of physician consensus to obtain the reference standard for the contextual feature values. A second physician board-certified in internal medicine independently assigned values to the three contextual properties for every clinical condition annotated by the first physician. JND looked at the disagreements on the assignments and changed his original values on cases for which he believed the second physician to be correct.

To evaluate agreement between the two physician annotators on the values they assigned to the contextual features in the test set before they came to consensus, we calculated for each contextual property both observed agreement (number of values both annotators agreed on divided by total number of annotations) and Cohen's kappa, which adjusts for chance agreement and differing frequencies of features. We also calculated the positive specific agreement score (agreement on the subset of annotations for which one or both annotators changed the default value of a feature) and negative specific agreement score (agreement on the subset of annotations for which one or both annotators kept the default value of a feature). These scores will be discussed in Section 4.

3.4. Development and evaluation

In this paper we evaluate ConText's ability to identify the values for the contextual properties negation, temporality, and experiencer for clinical reports of six different types. As described in the previous section, we split our corpus of clinical reports into a development set and a test set. Since ConText was originally developed for only one type of report, namely emergency department reports, we used the development set (containing 20 reports of each type) to port ConText to the five other report types in our dataset. We processed the development set with the original version of ConText and calculated the outcome measures described below. We performed a detailed error analysis, examining false negative and false positive classifications of contextual properties for each report type. Based on the error analysis, we made several changes to ConText. For each change, we reran a version of ConText system with the implemented change over the development set to assess its effect on performance. Most changes affected more than one report type and improved performance.

In keeping with the objective of this study –assessing the versatility of a simple, trigger term-based approach for detecting various contextual properties across several report types– we did not make any changes to the underlying algorithm; we used the results of the error analysis to update the lists of trigger terms, pseudo-trigger terms, and termination terms used by the algorithm. The application of this updated version of ConText to the independent, blind test set of 120 reports forms the basis for the results and discussion in Sections 4 and 5. In the remainder of this section we will briefly discuss the updates we made to ConText based on the error analysis on the development set.

The error analysis of the 120 documents in the development set yielded around 10 additional trigger terms, pseudo-trigger terms, and termination terms to be included in ConText's term lists. Some of the new terms are specific to the new report types included in our dataset, such as the pseudo-triggers "clinical history" for surgical pathology reports ("history" in this phrase should not trigger the value *historical* for temporality – although "clinical history" refers to a condition occurring sometime in the past, the condition is generally recent enough to merit the radiology exam and is therefore not a historical condition), and "without contrast" in radiology reports ("without" in this phrase should not trigger the value *negated* for the property negation because it modifies procedures rather than conditions). However, most additions were of a general nature, e.g., the trigger terms "roommate" for experiencer value *other* and "status post" for temporality value *historical*. Note that ConText does not keep separate term lists for each type of report.

We also made three changes pertaining to the application of termination terms to contextual properties. For instance, the Etiology class of termination was added to the set of termination term classes that apply to the contextual property negation.

Finally, the error analysis showed that ConText performed badly on chronic conditions and risk factors, i.e., alcohol, drug, and tobacco use and allergies. These conditions are generally considered historical in the reference standard, but almost always appear without explicit trigger terms in the text, whence ConText failed to change the default value *recent* to *historical* for these conditions. To address this issue we set the default value for the contextual feature temporality for chronic conditions and risk factors to *historical*. This change was implemented for a list of about 25 conditions determined by our medical expert to be chronic in nature, as well as the risk factors allergies, alcohol use, drug use, and smoking.

3.5. Outcome measures

For each report type, we assessed ConText's performance on each contextual property (negation, temporality, and experiencer). A true positive was counted when ConText correctly changed the default value of a contextual property; a true negative when ConText correctly did not change the default value; a false positive when ConText incorrectly changed the default value; and a false negative when ConText incorrectly did not change the default value. For example, changing the value of temporality for a condition from default *recent* to *historical* when the reference standard annotation for the condition is *recent* is a false positive.

From the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), we calculated recall (TP/(TP + FN)) and precision (TP/(TP + FP)) and their 95% confidence intervals, using calculations described by Newcombe [32]. We also calculated the F-score using equal weights for recall and precision. To explore differences in the report types, we examined the trigger terms most frequently applied to each report type.

Note that in the implementation of the ConText algorithm that we used for this study *historical* and *hypothetical* are both values of the contextual property temporality. However, as each value has its own set of trigger terms and termination terms, we will treat these values as if they represent two separate contextual properties called historical and hypothetical when discussing the evaluation results.

4. Results

In this section we present the results and outcomes of our study regarding the annotation of the reference standard, the scope experiment, and ConText's performance on the test set.

4.1. Reference standard

The reference standard physician generated 4654 annotations in the combined development and test set of 240 reports: 2377 annotated conditions in the development set and 2277 annotated conditions in the test set. Table 2 shows for each report type in the test set the mean number of annotated conditions per report, the mean number of sentences per report, and the mean sentence length per report. We also computed these statistics for the reports in the development set and found that the development set and the test set are very similar except for statistically significantly differences regarding the mean number of words in surgical pathology reports and discharge summaries, and the mean sentence length in discharge summaries (Student's t-test result with a $P < .05$, [33]).

The numbers in Table 2 illustrate some textual differences between the six types of reports. Surgical pathology reports, discharge summaries and emergency department reports are relatively long in terms of number of words per report. Echocardiogram reports show the least variation with regard to the number of words per report because these reports involve discussions of relatively standard procedures with a restricted set of findings and outcomes, whereas the other types of report deal with medical situations that are much more diverse in character.

The various types of reports are more similar with regard to average sentence length, with the exception of surgical pathology reports. This type of report tends to contain fairly long stock phrases describing test protocols and FDA approvals. Additionally, the other reports generally contain summary lists of medications, findings and diagnoses, which increases the number of short sentences found in these reports.

There is a large variation between report types concerning the number of conditions per report, ranging from around 45 for emergency department reports to around three for surgical pathology reports. About one third of the surgical pathology reports in the test set describe molecular diagnostic results and do not contain any clinical conditions at all, explaining the relative large spread in number of conditions per report for this type of report. The set of radiology reports contains an outlier mentioning 36 conditions, causing a relatively large spread in number of conditions per report for this report type as well. Without this outlier the standard deviation is 5.2.

Even though there is variation among the report types in terms of the textual features given in Table 2, this kind of variation is not expected to have a measurable effect on the performance of ConText. ConText's token-level regular expressions are applicable regardless of sentence length etc. In this respect, ConText's linguistic simplicity may be a strength when it comes to the algorithm's portability across different report types.

Table 3 shows the prevalence of the values of the three contextual properties negation, temporality, and experiencer across the six report types in the reference standard for the test set. This table will be discussed in more detail in Section 4.

Table 4 provides an overview of the trigger terms occurring in the test set. It should be noted that, since trigger terms are not annotated in the reference standard, the terms given in Table 4 are the ones found by the ConText system, and so this set includes trigger terms leading to false positives and excludes trigger terms leading to false negatives. The figures indicate that emergency department reports and discharge summaries generally have the most diverse sets of trigger terms. Also, especially for report types that contain many trigger terms, the frequency distribution of the terms is not uniform. For example, in emergency department reports, the trigger term "no" for *negated* occurs about three times as often as forms of the verb "to deny," which is the next most frequent trigger term. Not surprisingly, "history" is the most frequent trigger term for *historical* in all report types. "No" tops the list for *negated*, except in operative procedure notes and surgical pathology reports, reflecting a difference among report types in choice of words to convey negation.

Table 5 shows observed agreement and kappa coefficients, split out by report type, for the two physicians before they came to consensus on the contextual values for the annotated conditions in the test set. Table 6 provides positive and negative specific agreement scores for each of the contextual properties over all conditions in the test set.

For all contextual properties in Table 5 observed agreement is high, due to frequent agreement on the "negative" cases (i.e., where the default value of a contextual property is left unchanged), as indicated by the high negative specific agreement scores in Table 6, combined with the fact that there are many more negative cases than positive cases (cf. Table 3). Kappa values are reasonable for negation (.74) and experiencer (.67), but cause concern for hypothetical (.55) and historical (.35). The specific agreement scores show that the annotators agree more on when to keep the default value of a contextual feature than on when to change it.

Table 2
Average number of words per report, average sentence length in words per report, average number of annotated clinical conditions per report and standard deviations (SD) for each of the six report types in the test set. Note that there are 20 reports of each type in the test set. The last row in the table shows the averages and standard deviations calculated over the set of all report types in the test set. Numbers in bold represent the minimum value in a column; underlined numbers the maximum value in a column.

Report type	# of words		Sentence length		# of conditions	
	Mean	SD	Mean	SD	Mean	SD
Surgical pathology	790	480	<u>22.1</u>	16.9	**3.5**	3.8
Operative procedure	**613**	106	15.6	11.0	8.8	5.3
Radiology	<u>2467</u>	156	13.0	7.2	7.5	8.4
Echocardiogram	565	79	**11.2**	8.6	28.8	6.6
Discharge summary	856	390	15.8	12.7	20.9	16.1
Emergency department	913	301	12.9	9.7	<u>44.6</u>	20.5
All	664	364	14.6	11.8	19.0	18.4

Table 3
Overview of the prevalence of the values *affirmed*, *negated*, *recent*, *historical*, *hypothetical*, *patient*, and *other* for the contextual properties Negation, Temporality, and Experiencer for all annotated clinical conditions occurring in each of the six report types (SP: surgical pathology, DS: discharge summary, ECHO: echocardiogram, ED: emergency department, OP: operative procedure, and RAD: radiology) in the reference standard for the test set. The column "Total" indicates the total number of conditions in the set of all reports of a given type. Each cell provides the absolute number (proportion) of clinical conditions found in the set of reports of a given type that have the specified contextual property value.

	Total	Negation		Temporality			Experiencer	
		Affirmed	*Negated*	*Recent*	*Historical*	*Hypothetical*	*Patient*	*Other*
SP	69	65 (94%)	4 (6%)	65 (94%)	3 (4%)	1 (1%)	69 (100%)	0 (0%)
OP	175	158 (90%)	17 (10%)	171 (98%)	3 (2%)	1 (1%)	175 (100%)	0 (0%)
RAD	150	115 (77%)	35 (23%)	148 (99%)	2 (1%)	0 (0%)	150 (100%)	0 (0%)
ECHO	575	540 (94%)	35 (6%)	575 (100%)	0 (0%)	0 (0%)	575 (100%)	0 (0%)
DS	417	343 (82%)	74 (18%)	331 (79%)	60 (14%)	26 (6%)	416 (100%)	1 (0%)
ED	891	566 (64%)	325(36%)	778 (87%)	93 (10%)	20 (2%)	887 (100%)	4 (0%)
All	2277	1787 (78%)	490 (22%)	2068 (91%)	161 (7%)	48 (2%)	2272 (100%)	5 (0%)

Table 4
Three most frequent trigger terms and number of unique trigger terms for the contextual property values *negated*, *historical*, *hypothetical*, and *other* for each of the six report types (SP: surgical pathology, DS: discharge summary, ECHO: echocardiogram, ED: emergency department, OP: operative procedure, and RAD: radiology) and all reports combined in the test set as applied by the ConText algorithm in the test set. Each cell contains up to three trigger terms, each of which is followed by the number of instances in which that trigger term was used by ConText to change the default value of a contextual property to the value given in the first row. The number of unique trigger terms (*Unique)* refers to the total number of different trigger terms found in the test set. For example, for operative procedure reports we list the three most frequent trigger terms for the value *negated* ("no," "no evidence of," and "free"), but two other trigger terms (namely, "not" and "with no") were also used for a total of five unique trigger terms.

	Negated		*Historical*		*Hypothetical*		*Other*	
SP	No evidence of	2	History	4				
	No	1						
	Ruled out	1						
	Unique:	*3*	*Unique:*	*1*	*Unique:*	*0*	*Unique:*	*0*
OP	No evidence of	12	For approximately four months					
	No	4		2				
	Free	2	Status post	1				
	Unique:	*5*	*Unique:*	*2*	*Unique:*	*0*	*Unique:*	*0*
RAD	No	20	History	3				
	No evidence of	5						
	No significant	1						
	Unique:	*3*	*Unique:*	*1*	*Unique:*	*0*	*Unique:*	*0*
ECHO	No	15						
	Without evidence of	6						
	No evidence of	1						
	Unique:	*4*	*Unique:*	*0*	*Unique:*	*0*	*Unique:*	*0*
DS	No	22	History	34	If	17	Family history	1
	Denied/denies	16	Experiences	17	Return	6		
	Without	11	Status post	2	Should the	1		
	Unique:	*11*	*Unique:*	*6*	*Unique:*	*3*	*Unique:*	*1*
ED	No	176	History	56	Return	14		
	Denied/denies	59	For the last	2				
	Not	20	several weeks Prior	1				
	Unique:	*18*	*Unique:*	*3*	*Unique:*	*1*	*Unique:*	*0*
All	No	248	History	97	Return	20	Family history	1
	Denied/denies	76	Experiences	17	If	17		
	No evidence of	28	Status post	3	Should the	1		
	Unique:	*18*	*Unique:*	*8*	*Unique:*	*3*	*Unique:*	*1*

Table 5
Agreement between physicians before coming to consensus for the contextual value annotations for the properties Negation, Historical, Hypothetical, and Experiencer in each report type (SP: surgical pathology, OP: operative procedure, RAD: radiology, ECHO: echocardiogram, DS: discharge summary, ED: emergency department), as well as all reports combined (All) in the test set: Observed Agreement (A_o) and Cohen's Kappa (K). Also given is the number of conditions in the reference standard for which the given contextual property is assigned a non-default value (N; copied from Table 3). Empty cells indicate situations for which Kappa is undefined.

	Negation			Historical			Hypothetical			Experiencer		
	A_o	K	N	A_o	K	N	A_o	K	N	A_o	K	N
SP	.90	.35	4	.93	.30	3	1.0		1	1.0		0
OP	.96	.72	17	.98	.0	3	.99	.0	1	1.0		0
RAD	.82	.45	35	.97	.0	2	.99	.0	0	1.0		0
ECHO	.98	.80	35	1.0		0	.99	.0	0	1.0		0
DS	.88	.57	74	.88	.29	60	.96	.68	26	.99	.0	1
ED	.89	.77	325	.92	.38	93	.97	.52	20	1.0	1.0	4
All	.91	.74	490	.94	.35	161	.98	.55	48	1.0	.67	5

Table 6
Positive Specific Agreement (PSA) and negative specific agreement (NSA) between physicians before coming to consensus for the contextual property annotations for all reports in the test set.

	Negation	Historical	Hypothetical	Experiencer
PSA	.79	.37	.56	.67
NSA	.94	.97	.99	1.0

4.2. Scope of trigger terms

As discussed in Section 3.1.5, in NegEx the scope of a trigger term can never extend beyond a window of six tokens following a trigger term, whereas in ConText the scope runs until the end of the sentence or stops at a termination term, whichever comes first. Experimental results show that extending the scope in this way improves performance. We compared the performance of the current version of ConText with a version of ConText that includes NegEx's definition of scope. On our development set, extending the scope leads to a small increase in the number of false positives, as some conditions which are outside the six-token window will now, incorrectly, be assigned a non-default value for one or more of the contextual properties in cases where there is no intervening termination term between the trigger term and the condition. However, this increase in false positives is more than offset by an increase in true positives as a result of the converse effect: conditions outside the six-token window that now correctly are assigned non-default values. Overall, as shown in Table 7, precision stays the same or goes down, and recall and F-measure stay

Table 7
Recall, Precision, and F-measure for ConText with alternative definitions of scope for the contextual properties *negated*, *historical*, *hypothetical*, and *other*, evaluated for the 2377 annotated conditions in the development set. *N* is the number (proportion) of conditions in the reference standard for which a contextual property was assigned a non-default value. The two definitions of scope are: six-token window (''stw") and end-of-sentence (''eos"). The highest score for a given measure, report type, and contextual property is in bold.

	N (%)	Recall		Precision		F-measure	
		stw	eos	stw	eos	stw	eos
Development set: 2377 annotated conditions							
Negated	491/21	.96	**.98**	**.99**	.98	.97	**.98**
Historical	256/11	.70	**.79**	**.78**	.77	.74	**.78**
Hypothetical	56/2	.34	**.93**	**1.0**	**1.0**	.51	**.96**
Other	6/0	**.67**	**.67**	**1.0**	**1.0**	**.80**	**.80**

the same or go up. Comparison of the 95% confidence intervals calculated according to [32] confirms that there is a statistically significant difference in the recall scores between the two approaches for the contextual property hypothetical. The difference in recall for historical is on the border of being statistically significant.

As can be seen in Table 7, the differences are generally small, except for recall for the values *historical* and *hypothetical*. Discharge and emergency reports in particular often contain enumerations of historical or hypothetical conditions, as for example in ''Past medical history is positive for hyperlipidemia, hypothyroidism, weakness and difficulty with ambulation, aortic/mitral disease, spinal stenosis, cataracts, measles, mumps, and varicella in the past" and ''She will call Dr. N. at that same number if there is any newer increased shortness of breath, new onset of chest pain, lightheadedness, dizziness, fainting, ankle swelling, abdominal bloating/weight gain of 2 pounds in 24 hours or 4 pounds in a week or less." The tails of these lists fall outside the six-token window, depressing recall for this approach.

The slight drop in precision combined with the noticeable increase in recall shows that the combination of termination terms and sentence ends provides an effective mechanism for determining the scope of trigger terms. Based on the results of our scope experiment we decided to use ConText with this ''liberal" definition of scope for the other experiments described in this paper.

4.3. Performance on the test set

Tables 8a–d summarize ConText's performance on the conditions in the test set for each report type and contextual property as well as the number of cases for each situation. Shaded cells indicate confidence intervals that do not overlap with those of emergency department reports for which the algorithm was originally developed. The wide confidence intervals for some of the results reported in Table 8a–d (e.g., negation and historical in surgical pathology reports and experiencer in discharge summaries) are probably due to small sample sizes for these cases.

5. Discussion

In this section we will discuss the results of our study from various angles. The first two subsections concern the reference standard, focusing on inter-annotator agreement and the prevalence of the contextual properties across the various report types. The next subsection provides a comprehensive error analysis of ConText's results for the test set. This section closes with a discussion of the portability and transferability of the contextual properties negation, historical, and hypothetical.

Table 8
(a–d) Recall (R), precision (P), and F-measure (F), and 95% confidence intervals for recall and precision for the ConText algorithm on the test set for the contextual properties Negation, Historical, Hypothetical, and Experiencer. An empty Recall or Precision cell indicates that that no score could be calculated because the sum of true positives and false positives or the sum of true positives and false negatives is zero. The F-measure is provided only when both recall and precision are available. Recall and precision cells with confidence intervals that do not overlap with the confidence intervals for ED reports are shaded. To assess the significance of ConText's performance figures for a given contextual property, the number (proportion) of conditions in the reference standard for which the contextual property was assigned a non-default value is given in the final column (*N*; copied from Table 3).

	R	P	F	*N*
(a) Negation				
Surgical pathology	.75 .30–.95	.75 .30–.95	.75	4 6%
Operative procedure	.94 .73–.99	.84 .62–.94	.89	17 10%
Radiology	.86 .71–.94	1.0 .89–1.0	.93	35 23%
Echocardiogram	.91 .78–.97	.97 .85–.97	.94	35 6%
Discharge summary	.89 .79–.94	.84 .74–.90	.86	74 18%
Emergency department	.93 .90–.95	.96 .93–.98	.95	325 36%
All	.92 .89–.94	.94 .91–.96	.93	490 22%
(b) Historical				
Surgical pathology	.33 .06–.79	.17 .03–.56	.22	3 4%
Operative procedure	.0 .0–.56	.0 .0–.56		3 2%
Radiology	.0 .0–.66	.0 .0–.56		2 1%
Echocardiogram				0 0%
Discharge summary	.68 .56–.79	.77 .64–.87	.73	60 14%
Emergency department	.86 .78–.92	.82 .74–.89	.84	93 10%
All	.76 .69–.82	.75 .68–.81	.76	161 7%
(c) Hypothetical				
Surgical pathology	0.0 0.0–.79			1 1%
Operative procedure	0.0 0.0–.79			1 1%
Radiology				0 0%
Echocardiogram				0 0%
Discharge summary	.92 .76–.98	1.0 .86–1.0	.96	26 6%
Emergency department	.65 .43–.82	.93 .69–.99	.76	20 2%
All	.77 .63–.87	.97 .87–1.0	.86	48 2%
(d) Experiencer				
Surgical pathology				0 0%
Operative procedure				0 0%
Radiology				0 0%
Echocardiogram				0 0%
Discharge summary	1.0 .21–1.0	1.0 .21–1.0	1.0	1 0%
Emergency department	1.0 .51–1.0	1.0 .51–1.0	1.0	4 0%
All	1.0 .57–1.0	1.0 .57–1.0	1.0	5 0%

5.1. Reference standard

As shown in Table 4, the Kappa scores for the annotation of the contextual properties negation and experiencer in the reference standard are reasonable. A significant cause of disagreement between the annotators concerning negation was confusion about the proper annotation of internally negated phrases denoting conditions, such as ''afebrile" and ''unresponsive". Furthermore, the second annotator frequently missed negation signals such as ''denied," e.g., ''The patient denied headache," and ''resolved," e.g., ''His initial tachycardia resolved." The few disagreements for experiencer appear to be accidental mistakes (note that there are very few cases for which experiencer does not have the default value *patient*).

The second annotator used hypothetical sparingly, missing several typical cases such as ''Return to the emergency department if she develops abdominal pain." Furthermore, in disagreement with the second annotator, the first annotator annotated as hypothetical those diagnoses that were considered by but not fully committed to by the physician, as in, for example, ''This made us somewhat concerned about meningitis" and ''Early ARDS cannot be ruled out." Including annotations for certainty could increase agreement on hypothetical annotations.

For the contextual feature historical, chronic diseases and other long-term conditions such as asthma and hypertension constituted

a major source of disagreement between the two annotators. The first annotator marked these as *historical*, unless the conditions were clearly related to the reason for the patient's visit to the ED (for emergency department reports), their hospitalization (for discharge summaries), or procedures described in the other report types. The second annotator generally annotated chronic conditions as *recent*. Additionally, allergies and social histories, i.e., tobacco, drug, and alcohol use, were annotated as *historical* by the first annotator, and as *recent* by the second annotator.

A possible explanation for the low inter-annotator agreement scores is that the annotation task is difficult for humans to perform. In our situation this may be true to some extent for assigning temporality features to conditions: the distinction between historical and recent conditions is difficult to define in purely temporal terms. The original guidelines suggest putting the boundary between historical and recent at 2 weeks prior to the clinical visit or to procedures described in the report. However, in his temporality annotations, the first annotator also took into account the relevance of a condition to the current visit or procedure. For example, the mention "history of neck pain," i.e., neck pain which can be assumed to have started earlier than 2 weeks previous to the visit, in a radiology report for a spine MR was tagged as *recent* by the first annotator because the condition is the indication for the examination.

The major contributing factor to the depressed agreement figures, however, was the organization of the annotation process itself. Prior to annotation, both annotators were trained for the annotation task. For the second annotator there was a significant time lapse between the first set of annotations (for the 120 documents in the development set) and the second set of annotations (for the 120 documents in the test set). Between the two annotation periods, her command of the annotation guidelines decreased, resulting in a loss in of annotation quality for the second set. The second annotator processed the documents in one batch. We did not calculate intermediate agreement scores, which would have alerted us to any problems during the annotation phase. Because the low agreement scores are mostly a reflection of a suboptimal annotation process rather than the complexity of the annotation task, the physicians' agreement scores given in Tables 4 and 5 are not very informative in terms of providing an upper bound for ConText's performance reported in reported in Tables 8a–d.

Despite the issues with inter-annotator agreement discussed above, we believe that the reference standard itself is practically useful. By and large, the reference standard reflects the first, senior annotator's original annotations, with several changes based on feedback from the second annotator.

5.2. Prevalence of contextual properties across report types

Table 3 shows the distribution of the values of the three contextual properties negation, temporality, and experiencer for the conditions in the reference standard for the test set, split out by report type. The value *other* for experiencer is generally rare, as is *hypothetical* for temporality, except in discharge summaries. Discharge summaries often contain statements to the effect that the patient should return or call if a particular condition develops – these conditions are marked as hypothetical.

Historical conditions occur relatively frequently in emergency department reports and discharge summaries in comparison to the other report types. This is explained by the fact that emergency department reports and discharge summaries routinely include descriptions of the patient's past medical history.

Negated conditions can be found in all types of report, but their distribution varies widely among the six types. Negated conditions are most frequently encountered in emergency department reports.

The distribution of the contextual values in the development set is similar to the distribution observed in the test set, with the exception of historical events, which are more prevalent in the development set than in the test set. This is particularly true for surgical pathology reports. In the development set, 12 of the 55 conditions in surgical pathology reports were marked as historical (22%) vs. 3 out 69 (4%) in the test set. With an F-measure of .63, the updated version of ConText (see Section 3.4) did rather poorly on classifying the conditions in the surgical pathology reports in the development set as historical. To broaden the empirical basis of the discussion, we will include surgical pathology reports from the development set in the error analysis for the contextual property historical presented in the next section.

5.3. Error analysis

In order to assess the adequacy of a simple trigger term-based approach for contextual value assignment for different contextual properties and across different report types, we analyzed all the incorrect assignments, i.e., false positives and false negatives, that ConText produced for the set of conditions in the blind test set of 120 reports. The results of this error analysis are presented in Tables 9a–c. These tables only include data for contextual properties with more than 10 instances of non-default values in the test set (cf. Table 3). As explained above, concerning the contextual property historical for surgical pathology reports, we have combined the data from the development set and the test set for the error analysis. The contextual property experiencer is excluded from consideration altogether because of lack of sufficient data.

Each error made by ConText is assigned a "constructive" error category, reflecting the kind of change to the algorithm that is required to prevent the error. There are four error categories: the errors in the first category, "Missing terms," are false positives and false negatives that are caused by trigger terms, pseudo-trigger terms, and termination terms occurring in the text of the reports that are missing from ConText's term lists. Some missing terms we identified were "no longer present" (post-condition trigger term for negation), "history of recent" (pseudo-trigger term for historical) and "continues to have" (termination term for negation and hypothetical). The "Missing terms" class of errors also includes conditions missing from the list of chronic conditions that are assigned the value *historical* by default. All these errors can in principle be fixed by adding the missing terms to the appropriate term lists in ConText. We found nine missing terms for negation, seven for historical, two for hypothetical, and nine missing chronic conditions (note that there are several missing terms that are responsible for more than one error).

The category "Simple extensions" covers those errors for which there are no rules in the current version of ConText, but which can be addressed within the general framework of regular expressions and term lists. One simple extension is to introduce a class of terms that shelter conditions from the effect of a trigger term without terminating the scope of the trigger term. For example, possessive pronouns such as "his" and "her," and adjectives such as "unchanged" and "increased," when immediately preceding a condition, presuppose the existence of the condition. In a phrase like "A repeat of her CT head showed no acute stroke or bleed, unchanged residual right subdural hematoma, ...," the presence of "unchanged" neutralizes the effect of the negation trigger term "no" for the condition "right subdural hematoma." A rule to this effect can be implemented using the current mechanisms available to the ConText algorithm. Another extension implementable within the current framework is related to the rule that assigns the value *historical* to a list of chronic conditions. While this rule improves overall performance for the contextual property historical, it also introduces errors when the chronic condition occurs in

Table 9
(a–c) Classification of the errors made by the ConText algorithm for the contextual properties Negation, Historical, and Hypothetical for selected report types in the test set. The column "Total FP, FN" provides the total number of false positives and false negatives for each contextual property and report type. Four classes of errors are distinguished: Missing terms (error can be prevented by adding trigger terms, pseudo-trigger terms, or termination terms to ConText's term lists, or by extending ConText's list of chronic conditions), Simple extension (error can be prevented by adding another rule to the ConText algorithm), Outside framework (prevention of error requires a change that cannot be accommodated within the current ConText framework), and Annot./implem. (error is the result of a mistake in the reference standard annotations or a bug in the implementation of the ConText algorithm). The cells in the error classification columns give the absolute number and proportion of errors (with regard to the sum of false positive and false negatives) in each class for the given report type.

	Total FP, FN	Missing terms	Simple extension	Outside framework	Annot./implem.
(a) Negation					
Operative procedure	3 1	0 0%	0 0%	1 25%	3 75%
Radiology	0 5	0 0%	0 0%	0 0%	5 100%
Echocardiogram	1 3	4 100%	0 0%	0 0%	0 0%
Discharge summary	12 8	9 45%	4 20%	1 5%	6 30%
Emergency Department	13 22	3 9%	2 6%	9 26%	21 60%
All	29 39	16 24%	6 9%	11 16%	35 52%
(b) Historical					
Surgical pathology	6 8	2 14%	4 29%	6 43%	2 14%
Discharge summary	12 19	12 39%	9 29%	9 29%	1 3%
Emergency department	17 13	9 30%	2 7%	15 50%	4 13%
All	35 40	23 31%	15 20%	30 40%	7 9%
(c) Hypothetical					
Discharge summary	0 2	0 0%	0 0%	2 100%	0 0%
Emergency department	1 7	3 38%	0 0%	5 63%	0 0%
All	1 9	3 30%	0 0%	7 70%	0 0%

contexts that are clearly identifiable as recent, such as chief complaint, reason for admission, and discharge diagnosis sections. The "chronic" rule can be qualified so that it will leave the default value *recent* intact in these contexts. These two simple extensions cover most of the errors in this category.

The errors in the next category, "Outside framework," can only be fixed by incorporating structures and knowledge that cannot be represented with regular expressions and term lists: addressing these errors requires extensions to the algorithm that fall outside the current ConText framework. The errors in this class point to two kinds of knowledge that are necessary but absent from ConText. First, linguistic knowledge can provide constituent structure and phrase boundaries that are useful for delineating the scope of trigger terms in the absence of obvious termination terms. For example, in the sentence "left lower extremity pain with negative doppler," the post-condition negation trigger "negative" should not be able to reach outside the prepositional phrase and affect the condition "pain" inside the noun phrase modified by the prepositional phrase. Similarly, in the sentence "The patient denies tobacco use, drinks socially," the scope of the negation trigger "denies" should include only the complement of the verb "denies," i.e., the risk factor "tobacco use," and exclude the independent verb phrase "drinks socially" (which maps onto the risk factor "alcohol use"). In other cases, syntactic and semantic interpretation is necessary to identify relationships between conditions and (explicitly dated) events mentioned in the text. Consider, for example, the sentence "The patient is a 60-year-old male who recently had a significant history for coronary artery disease and states that he is being assessed for the possibility of a heart transplant and has refused coronary catheterization in the past due to anxiety." The condition "anxiety" in this sentence is *historical* because of its relationship with the procedure coronary catheterization, which was refused in the past.

The other crucial type of knowledge lacking from ConText is medical knowledge regarding conditions and their status and relationships within clinical reports. For example, a condition in an emergency department report may be accompanied by a trigger term that makes the condition historical, but because of its relationship with the chief complaint or current diagnosis, which are always recent, the condition itself should be considered recent as well, according to the expert annotations in our study. Thus, in a discharge summary listing "chronic pancreatitis" as one of the diagnoses, this condition should be marked as *recent* in the sentence "The patient is a 30-year-old woman with a history significant for chronic pancreatitis" despite the presence of the trigger "history" and the chronic nature of the condition.

The final source of errors distinguished in Table 9a–c are mistakes in the annotation of the reference standard, bugs in the implementation of the ConText algorithm, and erroneous output from supporting NLP components, in particular the sentence splitter and section identifier. Annotation mistakes include incorrect values for the contextual properties, as well as incorrectly annotated conditions. A rather frequently observed annotation mistake in operative procedure notes and emergency department reports is the inclusion of the trigger term within the span of an annotated condition, as a result of which ConText is unable to recognize these trigger terms and act on them.

Within each of the four error categories further subdivisions are possible, but these are beyond the scope of this paper. The data presented in Tables 8 and 9 will provide the basis for the discussion of the applicability of the trigger term approach across report types and contextual properties in the next section.

5.4. Transferability and portability

In this section we interpret the evaluation results to assess whether the approach to negation identification based on surface trigger terms can be transferred to other contextual properties, and whether this simple approach is portable across report types.

Given the prevalence data in Table 3, the discussion in this section will focus on the following contextual properties and report types: (1) negation across all report types except surgical pathology notes, (2) historical for emergency department reports and discharge summaries, and (3) hypothetical for emergency department reports and discharge summaries. The contextual property experiencer is excluded from the discussion, because there is very little data for this property.

5.4.1. Negation

The NegEx algorithm for detecting negated clinical conditions on which ConText is based was originally developed for discharge

summaries. In our evaluation we applied ConText to a test set containing five additional types of reports. Considering the results in Table 8a and the overlaps in confidence intervals reported there, we conclude that the ConText algorithm performs comparably well on all report types for the contextual property negation, apart from discharge summaries, for which the precision score is significantly lower. Most of the false positives contributing to the low precision score for discharge summaries are due to missing terms, in particular the pseudo-trigger "with/without," which would have avoided six false positives. Four of the remaining false positives can be fixed by introducing "neutralizing" terms as discussed in the previous subsection as one of the proposed simple extensions.

Overall, considering the entire report set, we conclude that decisions regarding the negation status of a condition generally do not involve medical knowledge, which is not available to ConText. Access to linguistic knowledge will improve performance by making the determination of the scope of a trigger term more precise, but using termination terms to demarcate scope appears to be adequate. Also, crucially, the lexical clues or trigger words for negation, when they occur in multiple report types, have the same interpretation across report types. Therefore, the use of a shared set of trigger words, provided the set is relatively complete, will produce acceptable results for the contextual property negation across all report types.

5.4.2. Historical

The performance results from Table 8 relevant to the discussion of the contextual property historical are summarized in Table 10. We notice that the F-scores for the contextual property historical are about 10 percentage points below the F-scores for negation, for both emergency department reports and discharge summaries. This observation, coupled with the fact from Table 9 that the percentage of "Outside framework" errors for historical is larger than that for negation for both report types, means that it is unlikely that the performance for historical can be increased to levels comparable to negation, even if we were able to address all the errors in the other categories, i.e., add the missing terms, make simple extensions to the algorithm, and fix the annotation mistakes and the bugs in the implementation. We therefore conclude that the approach employing regular expressions and terms lists does not transfer completely successfully from the contextual property negation to the contextual property historical. However, even though ConText fares worse on historical than on negation, the reported scores in the .73–.84 range are still useful for clinical NLP systems.

ConText cannot detect historical conditions as well as negated conditions because there are some significant differences in the way the two contextual properties are expressed in clinical text. First, whereas the trigger terms for negation can all be interpreted as denoting a negation of some kind, the word "history," which is the main trigger term for the value *historical* can have different word senses or interpretations, not all of which will make the associated condition historical. For instance, in the sentence "The patient presented with history suggestive of peritonsillar abscess," the word "history" refers to the patient's past medical history as a whole, and therefore, in this instance, should not be considered as a trigger term assigning the value *historical* to the condition "peritonsillar abscess." Similarly, in surgical pathology reports, most conditions following the trigger "history" were annotated by the physicians as *recent* (e.g., "History of polyps"). In these reports, the word "history" is generally used in the sense of "indication for the current procedure."

Second, whereas negation triggers can be interpreted locally within a sentence or phrase, the interpretation of historical triggers is sensitive to the section or broader context within the report in which they occur. For example, the term "status post" generally indicates a past procedure or condition, but if mentioned as part of a statement of diagnoses in a discharge summary, the condition modified by "status post" can become recent. The reverse effect of this dependence on the broader context is that a condition unaccompanied by a trigger term may nevertheless be historical. This happens, for example, when a physician begins describing the patient's past medical history and maintains this historical perspective for several sentences without using any explicit trigger terms. This issue reflects a fundamental difference of the notion of scope for negation and history. The scope of negation is grammatically restricted to a phrase and naturally bounded by a sentence, whereas the natural unit of scope for historical may be a discourse segment or sequence of sentences delineated by temporal expressions that switch the perspective between recent and historical.

Third, unlike negation, whether a condition is historical or recent can be contingent on its (temporal) relationship with other conditions or events in the report. In a sentence like "Results from a bone marrow biopsy (Apr 07, 06) raised the possibility of subtle involvement by a B-cell lymphoproliferative disorder," the disorder is historical because it was shown by a biopsy which was performed 2 years before the current visit. Also, conditions that are related to the chief complaint, reason for hospitalization, or current diagnosis are consistently recent, even if otherwise they would be considered historical. For example, in an emergency department report for a person coming in with a rash, the condition of being exposed to toiletries etc. in the phrase "no history of exposure to any new toiletries, soaps, or chemicals" will be recent, despite the trigger term "history." Note that connecting this condition to a rash requires medical background knowledge.

Finally, as alluded to in the last example, medical knowledge and reasoning may be necessary to decide whether a condition is recent or historical. Medical knowledge is in the general case irrelevant for negation (internally negated terms such as "afebrile," meaning "no fever" may be the exception). Consider for example the sentence "She takes oral contraceptive pills for ovarian cyst." This sentence contains no trigger term for historical, yet the condition ovarian cyst is marked as historical. One can reason that since contraceptive pills are usually taken for an extended period of time, the indication for this medication is a chronic or recurrent problem. Moreover, from the rest of the report it follows that the cyst is unrelated to current visit. Therefore this condition is historical (note that the "clinical act" described in this sentence is an observation by a physician, rather than the physician prescribing a medication; in the case of prescribing a medication, the cyst may be considered recent).

With regard to the question of portability of historical across report types, the overlapping confidence intervals in Table 8b show that the performance of ConText is similar for emergency depart-

Table 10
Summary of recall and precision scores (R, P), and F-measure (F) for the ConText algorithm on the contextual properties negation, historical, and hypothetical for the clinical conditions in emergency department reports and discharge summaries in the test set. The numbers are taken from Table 8.

	Negation			Historical			Hypothetical		
	R	P	F	R	P	F	R	P	F
Emergency department	.93	.96	.95	.86	.82	.84	.65	.93	.76
Discharge summaries	.89	.84	.86	.68	.77	.73	.92	1.0	.96

ment reports and discharge summaries. It appears emergency department reports and discharge summaries differ very little in the way historical events are expressed. Discharge summaries contain more chronic or pre-existing conditions, which are marked as historical. Otherwise there is no clear evidence in our report set for systematic differences between the two report types in this respect. As observed above, in both report types the general context in the report in which a condition occurs and its relation to other conditions mentioned in the report is very informative as to whether the condition is recent or historical.

5.4.3. Hypothetical

The results for the contextual property hypothetical shown in Table 10 look mixed at first glance. For emergency department reports, the F score is lower for hypothetical than for negation, whereas for discharge summaries the situation is reversed. Inspection of ConText's output reveals that four of the five errors in the "Outside framework" category for hypotheticals in emergency department reports (see Table 9) are false negatives related to diagnoses or concerns that originate from the patient rather than from the dictating physician, like "droopy" in "The patient states that she was concerned it may have been 'droopy.'" Conditions of this kind are annotated as hypothetical in the reference standard, because it is important to distinguish suspected diagnoses reported by a patient from diagnoses made by the physician. The distinction between suspected diagnoses reported by a patient (hypothetical) and symptoms reported by a patient (not hypothetical) is rather subtle and involves medical interpretation, and therefore will be hard to make automatically using a ConText-like approach. In our data set patient-offered diagnoses are more prevalent in emergency department reports than in discharge summaries – there is only one similar patient-reported diagnosis in our set of discharge summaries. Since in emergency department reports there is more emphasis on input from the patient than in discharge summaries, this asymmetry is probably a general trend.

Overall, given the performance figures in Table 10, we conclude that ConText produces results ranging from fair to very good for the contextual feature hypothetical. The approach based on regular expressions and trigger terms works well for identifying hypothetical conditions, except in the case of patient-offered diagnoses in emergency department reports. It is remarkable that just two trigger terms, "if" and "return" (see Table 4), cover most of the hypothetical conditions and generate very few false positives. The overlapping confidence intervals in Table 8 show that the approach works well for both emergency department reports and discharge summaries, although the presence of patient-offered diagnoses in emergency department reports lowers the recall scores for this type of report.

6. Future work

We have incorporated ConText in a pipeline-based NLP system called Topaz [34]. In spite of imperfect performance on historical classification, we still plan to use ConText to assign values to contextual properties of clinical conditions. We will update ConText based on our error analysis and plan an open source release of ConText as a stand-alone application and as an integrated module within Topaz.

Although ConText is extremely useful for assigning contextual features, its performance on identifying historical conditions is not completely satisfactory. We are therefore working on a more complex algorithm for determining whether a condition is historical. Based on results from this study, we believe that physicians use a variety of different types of information to determine whether a clinical condition is historical. In addition to local trigger terms, they consider medical knowledge about the condition itself, knowledge about the relationship between the condition and other events such as previous procedures or the current chief complaint, and knowledge about the context in which the condition occurred. Thus, we are exploring a more detailed annotation schema for temporality to try to model the information used by experts to infer historicity [35]. The schema includes explicit temporal expressions (as described by Zhou et al. [36]), related events, aspectual properties of a condition (e.g., if the condition is just beginning, whether it is continuing or resolved, etc.), and the context in which the condition is described in the exam. We are creating an automated classifier for determining whether a condition is historical based on input from the variety of features. In a statistical classifier (see Uzuner et al. [29] for a similar classifier), we can also consider report type as a feature, which may address report-specific usage of the same trigger term.

7. Conclusion

In this paper we have introduced an algorithm called ConText for determining whether clinical conditions mentioned in clinical reports are negated, hypothetical, historical, or experienced by someone other than the patient. ConText is based on the simple approach used by NegEx for finding negated conditions in text. The evaluation of ConText presented in this paper focused on two aspects: transferability of the trigger-term-based approach across different contextual properties and portability of the approach among different report types. Our results indicate that the approach used by ConText transfers well for identifying negated, hypothetical, and non-patient experiencers in all report types that demonstrate these properties. However, our study revealed an interesting distinction in the character of the contextual properties negation, hypothetical, and experiencer on the one hand and the contextual property historical on the other. Trigger terms for the non-default value *historical* are lexically ambiguous: the interpretation of some frequently occurring trigger terms is dependent on the type of report and the context within the report. Moreover, for some report types, knowledge about the clinical nature of a condition, the context in which it is mentioned in a report, and its relation to other conditions and events appearing in the report may be required to determine the proper value for historical. These factors limit the effectiveness of ConText's trigger-term-based approach to detecting historical conditions. A comprehensive solution to the task of detecting historical conditions must take into account information in addition to just trigger terms.

In spite of imperfect performance, ConText is a simple and easy-to-implement algorithm that can be integrated with any information extraction system for clinical reports; we believe ConText can improve precision of information retrieval and information extraction from various types of clinical reports containing negated, historical, non-patient, and hypothetical conditions.

Acknowledgments

This research was funded by 1 R01 LM009427-01 "NLP Foundational Studies and Ontologies for Syndromic Surveillance from ED Reports" and 1 K22 LM008301-01 "Natural Language Processing for Respiratory Surveillance." We acknowledge David Chu, MS, for his help in developing ConText. We would like to thank the anonymous reviewers for their useful comments.

References

[1] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301–10.
[2] Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999;74(8):890–5.

[3] Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161–74.
[4] Christensen L, Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding system. Proc workshop on natural language processing in the biomedical domain 2002:29–36.
[5] Hahn U, Romacker M, Schulz S. MEDSYNDIKATE – a natural language system for the extraction of medical information from findings reports. Int J Med Inf 2002;67(1–3):63–74.
[6] Elkin PL, Brown SH, Balas A, Temesgen Z, Wahner-Roedler D, Froehling D, et al. Biosurveillance evaluation of SNOMED CT's terminology (BEST Trial): coverage of chief complaints. Stud Health Technol Inform 2008;136:797–802.
[7] Taira RK, Soderland SG. A statistical natural language processor for medical reports. Proc AMIA Symp 1999:970–4.
[8] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17–21.
[9] Sneiderman CA, Rindflesch TC, Aronson AR. Finding the findings: identification of findings in medical literature using restricted natural language processing. Proc AMIA Annu Fall Symp 1996:239–43.
[10] Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. Medinfo 2004;2004:487–91.
[11] Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: a system for detecting and classifying encounter-based clinical events in any electronic medical record. J Am Med Inform Assoc 2005;12(5):517–29.
[12] Light M, Qiu XY, Srinivasa P. The language of bioscience: facts, speculations, and statements in between. BioLink 2004 workshop on linking biological literature, ontologies, and databases: tools for users. MA, Boston: 2004.
[13] Mizuta Y, Korhonen A, Mullen T, Collier N. Zone analysis in biology articles as a basis for information extraction. Int J Med Inform 2006;75(6):468–87.
[14] Medlock B. Exploring hedge identification in biomedical literature. J Biomed Inform 2008;41(4):636–54.
[15] Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma HB, et al. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc 2008;6:172–6.
[16] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. Evaluation of negation phrases in narrative clinical reports. Proc AMIA Symp 2001:105–9.
[17] Chu D, Dowling JN, Chapman WW. Evaluating the effectiveness of four contextual features in classifying annotated clinical conditions in emergency department reports. AMIA Annu Symp Proc 2006:141–5.
[18] Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. J Am Med Inform Assoc 2001;8(6):598–609.
[19] Elkin PL, Brown SH, Bauer BA, Husser CS, Carruth W, Bergstrom LR, et al. A controlled trial of automated classification of negation from clinical notes. BMC Med Inform Decis Mak 2005;5(1):13.
[20] Huang Y, Lowe HJ. A novel hybrid approach to automated negation detection in clinical radiology reports. J Am Med Inform Assoc. 2007.
[21] Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J Biomed Inform 2006;39(6):589–99.
[22] South BR, Chapman WW, Delisle S, Shen S, Kalp E, Perl T, et al. Optimizing syndromic surveillance text classifiers for influenza-like illness: does document source matter? Proc 2008 AMIA Fall Symp (under review), 2008.
[23] Mitchell KJ, Becich MJ, Berman JJ, Chapman WW, Gilbertson J, Gupta D, et al. Implementation and evaluation of a negation tagger in a pipeline-based system for information extraction from pathology reports. Medinfo 2004;2004:663–7.
[24] Aronson AR, Bodenreider O, Demner-Fushman D, Fung KW, Lee VK, Mork JG, et al. From indexing the biomedical literature to coding clinical text: experience with MTI and machine learning approaches. BioNLP: biological, translational, and clinical language processing. Prague: Association for Computational Linguistics; 2007. p. 105–12.
[25] Chapman WW, Cooper GF, Hanbury P, Chapman BE, Harrison LH, Wagner MM. Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders. J Am Med Inform Assoc 2003;10(5):494–503.
[26] Chapman WW, Chu D, Dowling JN. ConText: An algorithm for identifying contextual features from clinical text. BioNLP workshop of the association for computational linguistics, June 2007, Prague, Czech Republic: 2007. p. 81–8.
[27] Aronow DB, Feng F, Croft WB. Ad hoc classification of radiology reports. J Am Med Inform Assoc 1999;6(5):393–411.
[28] Gindle S, Kaiser K, Miksch S. Syntactical negation detection in clinical practice guidelines. In: Andersen SK, editor. eHealth beyond the horizon – get IT there. IOS Press; 2008. p. 187–92.
[29] Uzuner O, Zhang X, Sibanda T. Machine learning and rule-based approaches to assertion classification. J Am Med Inform Assoc 2009;16(1):109–15.
[30] Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: a framework and graphical development environment for robust NLP tools and applications. 40th anniversary meeting of the association for computational linguistics. Philadelphia, PA: 2002.
[31] Chapman WW, Dowling JN. Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. J Biomed Inform 2006;39(2):196–208.
[32] Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. Stat Med 1998;17(22):2635–50.
[33] Rosner B. Fundamentals of biostatistics. 3rd ed. Boston: PWS-Kent Publishing Co.; 1989.
[34] Chu D. Clinical feature extraction from emergency department reports for biosurveillance [Master's Thesis]. Pittsburgh: University of Pittsburgh; 2007.
[35] Mowery DL, Harkema H, Chapman WW. Temporal annotation of clinical text. BioNLP workshop 2008: current trends in biomedical natural language processing. Columbus, Ohio, USA; 2008. p.106–7.
[36] 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.