Automated Extraction of Substance Use Information from Clinical Texts

Yan Wang, MS[1], Elizabeth S. Chen, PhD[4, 5], Serguei Pakhomov, PhD[1, 2]
Elliot Arsoniadis, MD[1, 3], Elizabeth W. Carter, MS[4], Elizabeth Lindemann[1],
Indra Neil Sarkar, PhD, MLIS[4, 6], Genevieve B. Melton, MD, PhD[1, 3]
[1]Institute for Health Informatics, [2]College of Pharmacy, and [3]Department of Surgery,
University of Minnesota, Minneapolis, MN; [4]Center for Clinical & Translational Science,
[5]Department of Medicine, [6]Department of Microbiology & Molecular Genetics,
University of Vermont, Burlington, VT

Abstract

Within clinical discourse, social history (SH) includes important information about substance use (alcohol, drug, and nicotine use) as key risk factors for disease, disability, and mortality. In this study, we developed and evaluated a natural language processing (NLP) system for automated detection of substance use statements and extraction of substance use attributes (e.g., temporal and status) based on Stanford Typed Dependencies. The developed NLP system leveraged linguistic resources and domain knowledge from a multi-site social history study, Propbank and the MiPACQ corpus. The system attained F-scores of 89.8, 84.6 and 89.4 respectively for alcohol, drug, and nicotine use statement detection, as well as average F-scores of 82.1, 90.3, 80.8, 88.7, 96.6, and 74.5 respectively for extraction of attributes. Our results suggest that NLP systems can achieve good performance when augmented with linguistic resources and domain knowledge when applied to a wide breadth of substance use free text clinical notes.

Introduction

Social history (SH) factors including alcohol, drug, and nicotine use (collectively referred to as "substance use") are increasingly recognized as risk factors for preventable disease, disability, and mortality. A number of studies have been published describing the linkage between social risk factors and their associated morbidity or mortality(1-4). For example, nicotine abuse continues to be the leading preventable cause of morbidity and mortality in the United States(5), and the severity of substance use disorders is strongly associated with the magnitude of comorbidity(3). In 2006, the Institute of Medicine (IOM) report on "Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate" described the need for improving existing datasets, developing new data sources, and establishing strategies and models for incorporating behavioral and environmental factors and their interactions(6). In realizing this important link, the last decade has had increasing attention on substance use in its many forms with respect to health and disease(6-11).

The availability of electronic documents within electronic health record (EHR) systems provides an opportunity for clinical researchers to access a wide range of information about an individual's social environment and use of this information for secondary purposes (e.g., disease surveillance or evidence-based medicine) as well as primary uses for patient care (e.g., decision support). However, large amounts of detailed substance use information in EHR systems are stored predominantly in free-text rather than structured format(12, 13), underscoring the need for automated information extraction or other natural language processing (NLP) techniques specific for identifying social history information such as substance use.

In this study, we sought to develop an NLP system for detecting three main sub-categories of substance use (alcohol, drug and nicotine use) statements within free-text clinical notes and extracting related information (e.g., temporal, amount and type) within these substance use statements.

Background

Substance use information in clinical documents and structured modules in the EHR

This research group has previously performed a multi-institutional study on social history information in EHR system clinical notes(14). In this initial study, social history information contained in clinical notes from different sources (MTSamples(15), University of Vermont Medical Center [UVMMC; formerly Fletcher Allen Health Care], and University of Minnesota affiliated Fairview Health Services [FHS]) was analyzed and the adequacy of HL7 CDA-based models(16) and openEHR(17) archetypes for representing social history information within relevant statements across the three sources was studied. Table 1 shows an example of the alcohol use statement type identified in the study along with categories of information within statements for this type and a common set of data

elements and values. Table 2 illustrates the representation of information within a social history statement for nicotine use.

Among modules for EHR substance use, two follow-up studies showed that free-text comments with alcohol and nicotine use were often added due to the limited ability of structured parts of the EHR to describe some elements (e.g., amount and frequency) for both alcohol and nicotine usage (12, 13).

Table 1. Elements and values for alcohol use statement type.

SH statement type	Elements	Example value or pattern
Alcohol use	Status	current, past, nondrinker
	Temporal	[in/since/until] <date>
	Method	consume, use, drink
	Type	wine, alcohol, beer
	Amount	moderately, <#> [glasses/drinks/bottles/times]
	Frequency	occasionally, daily, rarely, on a weekly basis

Table 2. An **e**xample of representation of nicotine use statement.

"The patient has a history of heavy tobacco abuse for many years."	
Status = history	***Type*** = tobacco ***Temporal*** = many years
Amount = heavy	***Abuse*** = evidence of dependence

Extracting substance use information from clinical texts with NLP

Natural language processing (NLP) technologies have been used for extracting a wide range of information (e.g., drugs, diseases, and findings) from clinical notes(4, 18-20). Currently, most established clinical natural language processing systems (e.g., MedLEE(19) and cTAKES(21)) as well as clinical annotated corpus (e.g., MiPACQ(22)) are primarily focused on extracting named entities such as Unified Medical Language System (UMLS) concepts (e.g., diseases, medications, or procedures). A number of efforts have focused on using automated NLP techniques to extract smoking status (e.g., "Past smoker," "Current smoker," and "Smoker") and assess clinician adherence of tobacco treatment guideline from clinical notes such as discharge summaries(23-29). In Uzuner's work, the authors described several systems for classifying the smoking status of the patients by using machine learning and rule based algorithms. These systems presented in the paper reported F-scores from 84 – 90. In this study, our preliminary classification results showed that a rule-based classifier could achieve F-scores from 85.9 – 95.9. For extracting of deeper information related to substance usage other than status such as amount, type, there has been limited work on collecting these detailed information on substance use more comprehensively.

We previously incorporated a family history model into our open-source clinical NLP system, BioMedICUS(30), for automated extraction of family history information that identifies observations (e.g., disease or procedure), relative or side of family with the attributes (e.g., vital status, age of diagnosis, certainty, and negation) and predication. In this study, we sought to extend BioMedICUS for automated detection of substance use statements and extraction of detailed information within each statement. We aimed to utilize the patterns and lexicon collected from the multi-site sources used in previous work(14) along with the deep dependency relationships between tokens within statements provided by the Stanford Dependency parser(31) to extract elements from social history statements. To augment our system performance, we also leveraged semantic labels in the Propbank corpus and the MiPACQ corpus.

Methods

Corpora and annotation

The corpora used in this study included 491 "Consult - History and Physical" notes from MTSamples.com (MTS), a public web repository with about 5,000 sample clinical notes ("Development Corpus"), and 200 "HP" notes from the University of Pittsburgh Medical Center (UPMC) de-identified clinical notes repository(32) ("Evaluation Hold-out Corpus"). The final corpora included clinical notes in a range of different specialties from both acute and ambulatory settings. The corpora annotation consisted of three rounds of annotation. First, the collected clinical notes were annotated for the social history contents as well as social history section headers with the General Architecture for Text Engineering (GATE)(33), a Java framework for developing NLP pipelines. To establish

agreement regarding annotations, a subset of 100 notes were studied collectively by TJW and EWC prior to annotating the remaining set of annotations. In the next step, annotated social history text of each note was extracted from the GATE XML output and categorized into different sets of statement types (e.g., "ALCOHOL_USE," "OCCUPATION," and "DRUG_USE") with same annotation tool. Finally, sentences of three social history statement types, "ALCOHOL_USE," "DRUG_USE," and "TOBACCO_USE" were further annotated for elements (e.g., amount, frequency, and temporal) as well as the relationships between those elements with the brat rapid annotation tool (BRAT)(34), a web-based tool for structured annotation. Table 3 shows a brief description of the six elements that were extracted from our annotation guidelines for substance use.

Table 3. Description and examples of elements for alcohol, drug and nicotine use.

Element	Brief Description	Alcohol Use	Drug Use	Nicotine Use and Exposure
Amount	How much a substance used	• Few glasses wine/day • beer or wine 1-2 per month	• <1 joint QOD for pain • Daily 1pill 4X/day	• 1-2 cigarettes/day • heavy tobacco use
Status	Current or past substance use.	• quit many years ago • current drinker	• clean since 2004 • History of drug abuse	• Minimal smoker • Current tobacco
Type	Type of substance use or exposure	• Few glasses wine/day • beer or wine 1-2 per month	• Used OxyContin prescribed to mother • Overdose of Molly	• Occasional Cigar • Smoked pipes and cigars
Frequency	How often substance used	• one every 2 months • Occasional beer	• very rare • Marijuana 2x/week.	• 1-2 cigarettes/day • a pack per week
Method	How substance used or how exposed	• drinks wine	• Injected heroin for years • Occasional IVDU	• smokes a couple • secondhand smoke exposure
Temporal	Temporal information e.g., date started, age started, date quit and duration of use	• started drinking in 1985 • drink since age 20 • quit recently • drink for 20 years	• First joint in 1987 • methadone 2007 stopped • used 10 years	• started smoking in 1985 • quit years ago • Smoked for 4-5 yrs

Figure 1 shows the annotation of a set of alcohol use sentences using BRAT. For each social history statement type, 10% of the sentences of each statement type were annotated by two annotators for inter-rater agreement on the annotation of elements. Four informatics experts provided manual annotations, including a physician, one biomedical informatics PhD, and two biomedical informatics graduate students.

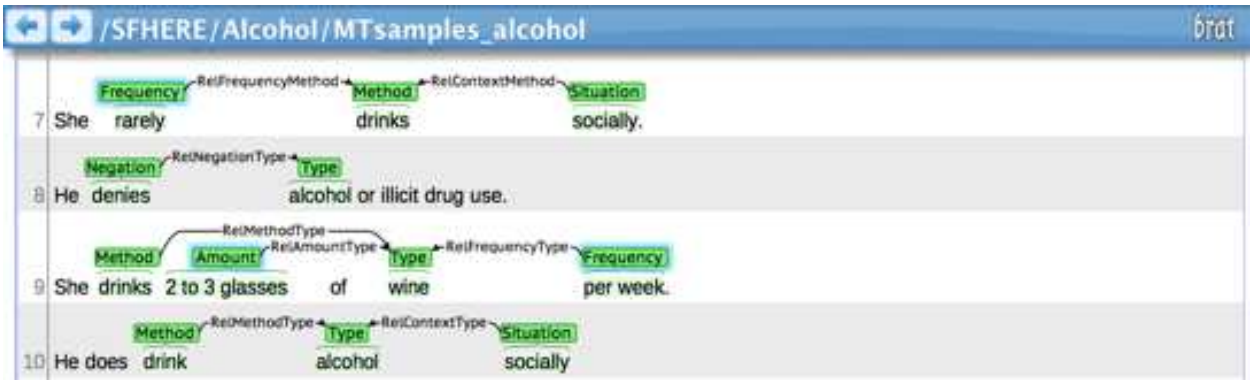

Figure 1. Annotation of "ALCOHOL_USE" sentences in clinical notes. Green shaded box represent elements and arrows represent relationships between elements.

Substance use statement detection

Figure 2 shows the overview of the substance use statement detection module development. The substance use statement detection rules were built with half of the 246 MTS "Consult - History and Physical" notes. Another two sets of MTS notes were used as a development set (N=123) and a holdout test set (N=122).

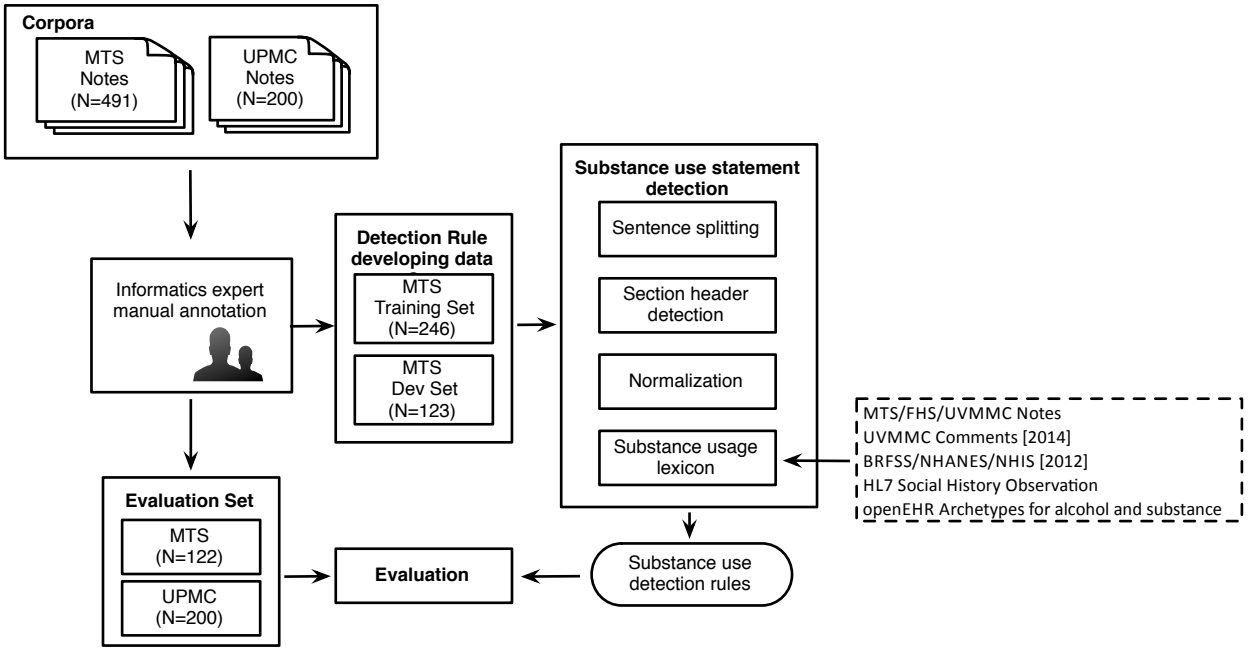

Figure 2. Overview of substance use statement detection rules developing.

All UPMC "H&P" notes were used as a separate evaluation set. Text was extracted and split into sentences by a sentence splitter. All tokens within were normalized based on the SPECIALIST Lexicon(35) and substance use sentences were extracted.

A substance use detection lexicon was collected from all our previous work on social history and substance use as key features for substance use statement detection (13, 14, 36). Overall, three lexicons for alcohol, drug, and nicotine were created from MTSamples/FHS/UVMMC notes, UVMMC comments, and HL7 Social History Observation and openEHR alcohol related archetypes. Section headers and sub-headers of each statement were also used as a feature for substance use statement detection. Section header identification was aided by a limited set of header expressions, capitalization, spacing, and punctuation header patterns. A set of regular expression based rules for substance use statement classification was manually developed using the lexicons, section headers, and iterative error analysis on the training dataset to improve the system performance. In each iteration, rules were modified and new rules were included to capture cues from section headers or sentence text. For example, if a sentence contains any family member, then the sentence was not classified as a substance usage statement. In total 10 rules were created for substance classification. The developed substance use statement detection system was evaluated on two evaluation sets: MTS test set notes (N=122) and UPMC evaluation notes (N=200).

In this study we also implemented a support vector machine (SVM) model for alcohol usage statement detection. Latent dirichlet allocation (LDA) and information gain (IG) were used for feature selection. Selected features are then used by supervised SVM machine learning to classify each statement in the training dataset. LDA with Gibbs Sampling (iteration = 1500) was implemented in Stanford Topic Model Toolkit (TMT-0.4.0)(37). Topic numbers range from 50–800 were chosen. Keywords for each topic were obtained from the output files. We then used IG to rank those keywords for each topic and chose filtered top numbers of keywords as features to implement classification.

Elements extraction

As shown in Table 4, the number of occurrences for some elements (e.g., amount and frequency) was limited in the small MTS development corpus. Thus, the attribute extraction module was developed on all MTSamples annotated

statements for alcohol (N=243), drug (N=130), and nicotine (N=311) use for the six elements (e.g., amount, frequency, and temporal). The UPMC annotated statements were used as the single holdout test set for evaluation of element extraction. Annotated sentences were first parsed by Stanford Parser(38) for constituent and dependency parses. The constituent parse of a sentence provides syntactic cues for elements extraction. For example, a phrasal phrase with a time period token (e.g., year and month) could be a temporal element (e.g., "since early this year"). Based on this fact, a set of rules was built to detect for temporal elements by using constituent parses. For instance, the constituent parses provided prepositional phrases which were helpful for temporal elements such that the pattern "(PP (IN (for|until|since|in|on|at)" and with a time marker could be used for temporal element detection. The dependency parse of a sentence can provide the dependency structure of the sentence. A dependency structure represents a directed graph between the tokens of a sentence, where edges denote pairwise grammatical relationships (e.g., determiner [the – patient] or adjectival modifier [significant - modified]).

Figure 3 shows an example dependency structure for "She denies any significant tobacco or alcohol..." The dependency structure of a sentence captures relationships between elements and the substance (e.g., alcohol or drug). These relationships can help to extract only related elements for a complicated sentence or sentences that with multiple substances. For instance, the dependency structure for "She is a smoker about a pack and a half for 38 years and notes rare alcohol use." can help to detect the right substance, smoking presumably tobacco or alcohol in this case, and the amount phrase "about a pack and a half for 38 years" is related to smoking. In our training dataset, we observed that for sentences like "No history of tobacco, alcohol, or illicit drug use.", which occurred frequently in clinical notes, the distance based algorithms is obviously will not be able to detect the correct relationship between entities. On the other hand, a dependency parse can easily detect such a relationship. In this study, we used Stanford Parser v3.5.1 to generate dependency structures for social history statements.

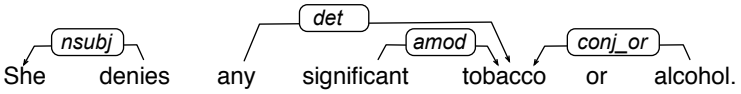

Figure 3. Dependency structure of sentence "*She denies any significant tobacco or alcohol ....*"

For temporal extraction, we observed that temporal elements varied largely in terms of syntactic patterns and lexicon (e.g., "*most of her adult life,*" "*up until about ten years ago,*" "*for a while*"). The syntactic patterns and lexicons collected from our previous studies were not adequate to extract temporal elements. In addition to the lexicon collected from previous study, we used the temporal semantic role (ARG-TMP) annotations from Propbank(39) and MiPACQ(40) – the latter of which is a clinical corpus annotated with PropBank-style predicate argument structures. The Propbank and MiPACQ are two large corpora annotated with semantic roles (e.g., goal, temporal, manner, purpose) of all the phrases in a sentence. This semantically annotated sentence shows an example of semantic role annotation on constituent parse which includes temporal expressions "TMP": " (S (NP-SBJ (DT The) (NN patient)) (VP (MD will) (VP (VB be) (ADJP-PRD (JJ ready) (NP-TMP (NNP Feb.) (CD 15))))) (. .)))". Temporal lexicons and patterns collected from the temporal semantic role annotations in these two resources were included into the rules for temporal elements detection. For example, temporal phrases like "at the same time" and "yesterday" were included as extra lexicon. The resulting new syntactic patterns captured additional temporal patterns including many that are not in the training dataset or the original lexicon and had improved performance. For each element of the three types of substance use statements, a set of regular expression heuristic rules were developed based on the lexicon, syntactic structure, and dependency structure. The created module was tested on substance use statements annotated from 200 UPMC notes.

In order to extract all 6 elements, each alcohol, drug and nicotine statement was first searched for patterns (e.g., "up to 6 drinks", "3-packs") and lexical items (e.g. "significant", "dips"), represented by regular expressions. Then, the constituent parse of the statement was searched for syntactic patterns of each element (e.g., "(PP (IN until)". The searched phrases were validated by dependency parse of each statement. The patterns, lexicon and syntactic patterns used for elements extraction were collected from multi-site sources in our previous work[14], MiPACQ and Propbank, with a number range from 6 to more than 3,000. A set of rules was manually created to determine if a searched phrase is related to the particular substance of interest by using iterative error analysis on the training dataset. In each iteration, rules were modified and new rules were included to accurately detect the correct relationships. Below are two examples of such rules. Some element types often needed more rules (e.g., 15 rules for temporal) to validate

which substance was associated with a searched phrase. On average, 5 rules (median 3, range 1-15) were created for each element extraction task.

Example 1). IF all parents of tokens in the searched phrase consist of the drug token AND the searched phrase is not negated THEN the searched phrase is associated with the drug token.

Example 2). IF the statement contains only the drug token without alcohol or nicotine keywords THEN the searched phrase is associated with the drug token.

Results

Table 4 summarizes the statistics of overall annotations of the corpora. From 491 MTS notes, 234 sentences were marked as alcohol use related, 124 were drug use related, and 260 were nicotine use related. From UPMC notes, 138, 72 and 148 sentences were marked as related to alcohol, drug and nicotine use. The number of occurrence of each element with these sentences is also listed in Table 4.

Table 4. Summary of corpora annotations.

(A) Substance use (SU) text annotation				
Source	**Note type**	**No. of notes**	**No. of notes with SU**	**No. of SU sentences**
MTS	Consult – H&P	491	378	1350
UPMC	H&P	200	179	651
(B) Substance use statement type annotation				
Source	**Statement type**	**No. of notes with statement type**		**No. of sentences**
MTS	Alcohol use	234		243
	Drug use	124		130
	Nicotine use	260		311
UPMC	Alcohol use	138		160
	Drug use	72		81
	Nicotine use	148		172

(C) Substance use elements annotation							
Source	**Statement type**	**No. of Amount occurrence**	**No. of Status occurrence**	**No. of Type occurrence**	**No. of Frequency occurrence**	**No. of Method occurrence**	**No. of Temporal occurrence**
MTS	Alcohol use	31	44	193	55	154	14
	Drug use	13	34	148	3	3	10
	Nicotine use	62	144	124	53	202	83
UPMC	Alcohol use	19	36	125	25	78	21
	Drug use	20	17	92	2	13	6
	Nicotine use	25	90	48	21	108	43

The Cohen's kappa inter-rater agreement between two annotators for elements of alcohol, drug, and nicotine use statements was 83.4, 80.9, and 90.6. Table 5 shows substance use statement detection performance on the MTS test set (N=122) and UPMC test set (N=200). As summarized, detection of alcohol, drug, and nicotine use statements was good with high F-scores.

Table 6 summarizes the performance of the system for extraction of substance use elements on UPMC substance use statements. The type, method and amount detection showed good performance based on F-scores for alcohol, drug and nicotine use sentences. Frequency and status detection for drug and nicotine use showed better performance for drug use statements and for nicotine use statements compared to alcohol use statements.

After inclusion of extra patterns and lexicon from general English semantic resources and clinical annotated semantic resources, temporal detection improved greatly, though this was still not as good as the performance of other elements, primarily due to the large variety of potential temporal expressions. Figure 4 shows the improvement of temporal element extraction with alcohol use and drug use statements achieved by adding in outside lexicons and patterns from MiPACQ and Propbank.

Table 5. Substance use statement detection performance.

	Sensitivity	Precision	F-score
Alcohol use			
MTS test set	91.8	97.8	94.7
UPMC test set	83.1	97.6	89.8
Drug use			
MTS test set	84.0	87.5	85.7
UPMC test set	76.0	98.3	85.7
Nicotine use			
MTS test set	96.6	95.0	95.8
UPMC test set	89.1	89.7	89.4

Table 6. NLP system element extraction performance for substance use statements.

Element	Sensitivity	Precision	F-score
Alcohol use			
Amount	68.0	94.4	79.1
Frequency	73.3	73.3	73.3
Status	75.0	61.4	67.5
Method	84.6	80.5	82.5
Type	98.4	87.2	92.5
Temporal	61.9	68.4	65.0
Drug use			
Amount	77.3	94.4	85.0
Frequency	100.0	100.0	100.0
Status	83.3	93.8	88.2
Method	76.9	100.0	87.0
Type	98.9	100.0	99.4
Temporal	66.7	100.0	80.0
Nicotine use			
Amount	84.0	80.8	82.4
Frequency	95.2	100.0	97.6
Status	76.7	100.0	86.8
Method	93.5	100.0	96.7
Type	95.8	100.0	97.9
Temporal	74.4	80.0	78.4

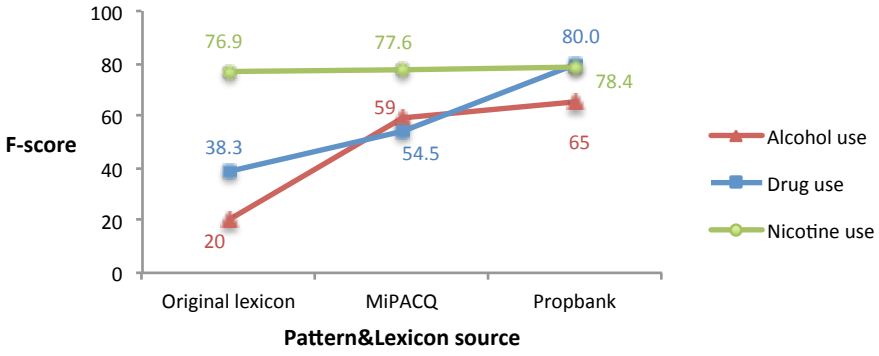

Figure 4. Temporal element detection performance improvement.

Discussion

In this study, we developed an NLP system for substance use statement detection and element extraction from clinical notes based on previous substance use models and the addition of developed lexicons(14). The system provides functionalities to detect three types of substance use statements: alcohol, drug, and nicotine use from free-text clinical notes, as well as extraction of important semantic elements as defined in substance use models previously developed(14) including amount, frequency, type, status, method and temporal. Overall, the developed system performed well, although certain elements like temporal expressions had challenges due to the variability in expression with these items.

Similar to previous work on family history extraction from clinical texts(41), we used the dependency structure for capturing relationships between phrases and tokens in the substance use statement. We found that the dependency relationship could help attaching possible element tokens or phrases to the right substance (e.g., alcohol or drug). However, due to dependency parsing errors on long, complicated sentences, relationships were sometimes wrongly collected from the dependency structure sometimes leading to errors on attaching tokens to correct types or methods. Further study is needed for reducing this type of errors. In our experiments, we did not observe sentences that caused the parser to break. Constituent parsing errors also caused sentence-segmenting errors, which in turn affected the element detecting process.

As shown in previous work(41), rule-based approaches can achieve comparable performance as machine learning approaches. In this study we choose to implement a rule-based NLP system by leveraging features from a number of sources and previous work, particularly to improve substance use lexicons. Substance use statement detection results of the rule-based systems showed good performance on both the MTS and UPMC test sets for all three types of substance use statements. Further, our pilot experiments on alcohol statement detection using models based on topic analysis and machine learning showed no performance improvement compared with the rule-based statement detection.

We recognize as a limitation that our system would ideally have a baseline gold standard with which to evaluate itself against. Currently, gold standard corpora and comparable systems do not exist. In the future, we plan to release publically available corpora with annotations so that future systems will have a baseline corpora with which to evaluate itself.

This study was also limited in that development of statement detection and elements extraction portions of the NLP system were built on one corpus (MTS notes) and tested on a relatively small hold-out corpus (UPMC notes). While this approach suggested that our findings could be generalized to clinical notes from other sites, validation with data from another institution would be helpful in confirming these results. A next step will therefore include setting further assessments of the system on institutional clinical notes and expansion of the system for substance use comments from the structured EHR social history module. Annotations used for extracting frequency, method and temporal for drug use statements were limited in this study. Although the same approach used for extracting these same elements for alcohol use and nicotine use had good results, validation with additional drug use statements would help with confirming our findings.

Conclusion

In this study, we achieved reasonable performance for both substance use statement detection and element extraction from two clinical notes corpora. The results of this study are promising for automated detection of free-text substance use statements and extraction of detailed substance use information from clinical notes. Next steps include further refinement of element extraction for better performance, additional evaluations to demonstrate generalizability of the system, and development of modules for extraction of additional social history statements.

Acknowledgements

The National Institutes of Health through the National Library of Medicine (R01LM011364 and R01GM102282), Clinical and Translational Science Award (8UL1TR000114-02) supported this work. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

References

1. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. *Jama*. 2004;291(10):1238-45.
2. Babor TF, Sciamanna CN, Pronk NP. Assessing multiple risk behaviors in primary care. Screening issues and related concepts. *Am J Prev Med*. 2004;27(2 Suppl):42-53.
3. Jane-Llopis E, Matytsina I. Mental health and alcohol, drugs and tobacco: a review of the comorbidity between mental disorders and the use of alcohol, tobacco and illicit drugs. *Drug Alcohol Rev*. 2006;25(6):515-36.
4. Huang FY, Ziedonis DM, Hu HM, Kline A. Using information technology to evaluate the detection of co-occurring substance use disorders amongst patients in a state mental health system: implications for co-occurring disorder state initiatives. *Community Ment Health J*. 2008;44(1):11-27.
5. Centers for Disease Control and Prevention. Smoking & Tobacco Use Fast Facts. [updated April 24, 2014]; Available from: http://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/.
6. In: Hernandez LM, Blazer DG, editors. Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate. Washington (DC)2006.
7. Ottman R. An epidemiologic approach to gene-environment interaction. *Genet Epidemiol*. 1990;7(3):177-85.
8. Ottman R. Gene-environment interaction: definitions and study designs. *Prev Med*. 1996 Nov-Dec;25(6):764-70.
9. Adler NE, Stead WW. Patients in context--EHR capture of social and behavioral determinants of health. *N Engl J Med*. 2015 Feb 19;372(8):698-701.
10. Primary Care and Public Health: Exploring Integration to Improve Population Health. Washington (DC)2012.
11. Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. Washington (DC)2015.
12. Chen E, Garcia-Webb M. An analysis of free-text alcohol use documentation in the electronic health record: early findings and implications. *Appl Clin Inform*. 2014;5(2):402-15.
13. Chen ES, Carter EW, Sarkar IN, Winden T, Melton GB. Examining the use, contents, and quality of free-text tobacco use documentation in the electronic health record. *AMIA 2014 Annual Symposium*. 2014:AMIA 2014 Annual Symposium.
14. Chen ES, Manaktala S, Sarkar IN, Melton GB. A multi-site content analysis of social history information in clinical notes. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*. 2011;2011:227-36.
15. MTSamples. [March 1, 2015]; Available from: http://www.mtsamples.com/.
16. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, et al. HL7 Clinical Document Architecture, Release 2. *J Am Med Inform Assoc*. 2006 Jan-Feb;13(1):30-9.
17. OpenEHR Specifications. [March 1, 2015]; Available from: http://www.openehr.org/svn/specification/TAGS/Release-1.0.2/publishing/architecture/rm/ehr_im.pdf.
18. Long W. Extracting diagnoses from discharge summaries. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*. 2005:470-4.
19. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods Inf Med*. 1998;37(1):1-7.
20. Pakhomov SV, Ruggieri A, Chute CG. Maximum entropy modeling for mining patient medication status from free text. *Proc AMIA Symp*. 2002:587-91.
21. cTAKES. Available from: http://ctakes.apache.org.
22. Albright D, Lanfranchi A, Fredriksen A, Styler WF, Warner C, Hwang JD, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*. 2013 January 25, 2013.
23. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*. 2008;15(1):14-24.
24. Liu M, Shah A, Jiang M, Peterson NB, Dai Q, Aldrich MC, et al. A study of transportability of an existing smoking status detection module across institutions. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*. 2012;2012:577-86.
25. McCormick PJ, Elhadad N, Stetson PD. Use of semantic features to classify patient smoking status. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*. 2008:450-4.
26. Wicentowski R, Sydes MR. Using implicit information to identify smoking status in smoke-blind medical discharge summaries. *J Am Med Inform Assoc*. 2008;15(1):29-31.
27. 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.

28. Hazlehurst B, Sittig DF, Stevens VJ, Smith KS, Hollis JF, Vogt TM, et al. Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines. *Am J Prev Med*. 2005;29(5):434-9.
29. Wu C-Y, Chang C-K, Robson D, Jackson R, Chen S-J, Hayes RD, et al. Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. *PLoS One*. 2013;8(9):e74262.
30. BioMedICUS (March 1, 2015).
31. de Marneffe M-C, MacCartney B, Manning CD. Generating Typed Dependency Parses from Phrase Structure Parses. . *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*. 2006:449-54.
32. University of Pittsburgh NLP Repository. Department of Biomedical Informatics, University of Pittsburgh; Available from: http://www.dbmi.pitt.edu/nlpfront.
33. GATE. [March 1, 2015]; Available from: https://gate.ac.uk.
34. Brat. [March 1, 2015]; Available from: http://brat.nlplab.org.
35. SPECIALIST Lexicon. Available from: http://www.nlm.nih.gov/pubs/factsheets/umls.html.
36. Melton GB, Manaktala S, Sarkar IN, Chen ES. Social and behavioral history information in public health datasets. *AMIA Annu Symp Proc*. 2012;2012:625-34.
37. Stanford Topic Model Toolkit. [March 1, 2015]; Available from: http://www-nlp.stanford.edu/software/tmt.
38. Stanford NLP tools. [March 1, 2015]; Available from: http://nlp.stanford.edu/software/index.shtml.
39. Palmer M, Gildea D, Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Comput Linguist*. 2005;31(1):71-106.
40. Cairns BL, Nielsen RD, Masanz JJ, Martin JH, Palmer MS, Ward WH, et al. The MiPACQ Clinical Question Answering System. *AMIA Annu Symp Proc*. 2011;2011:171-80.
41. Bill R, Pakhomov S, Chen ES, Winden TJ, Carter EW, Melton GB. Automated Extraction of Family History Information from Clinical Notes *AMIA annu Symp Proc 2014*. 2014:1709-17.