Vocabulary Development To Support Information Extraction of Substance Abuse from Psychiatry Notes

Sumithra Velupillai[1,2], **Danielle Mowery**[3], **Mike Conway**[3], **John Hurdle**[3] **and Brent Kious**[4]
[1] School of Computer Science and Communication, KTH, Stockholm
sumithra@kth.se
[2] IoPPN, King's College London
[3] Department of Biomedical Informatics, University of Utah
[4] Department of Psychiatry, University of Utah
firstname.lastname@utah.edu

Abstract

Extracting information from mental health records can be useful for large-scale clinical studies (e.g., to predict medication adherence or to understand medication effects) in this clinical specialty largely underserved by the Natural Language Processing (NLP) community. Vocabularies that contain medical terms for specific clinical use-cases, such as signs, symptoms, histories, social risk factors, are valuable resources for the development of NLP systems that aid clinicians in extracting information from text. Substance abuse is an important variable for many clinical use-cases, but, to our knowledge, there are no publicly available vocabularies that cover these types of terms. In this study, we apply and combine three methods for generating vocabularies related to substance abuse. We propose a simple and systematic method to generate highly relevant vocabularies and evaluate these vocabularies with respect to size and content, as well as coverage and relevance when applied to authentic psychiatric notes.

1 Introduction

Information about a mental health patient's clinical condition is documented routinely in mental health records, mostly in the form of free-text. Extracting information from these documents can be useful for large-scale clinical studies to develop new treatment alternatives, to understand disease progression and medication effects, etc. Vocabularies that contain relevant terms for specific clinical use-cases are useful resources for the development of Natural Language Processing (NLP) systems that aid clinicians in extracting information from text.

In this study, we focus on the problem of automated vocabulary generation, specifically, to automate the generation of relevant synonyms and related terms, focusing on *substance abuse*, an area not well-studied. Specifically, we aim to:

1. compare, assess, and combine three different automated vocabulary generation methods
2. determine vocabulary coverage and relevance in substance abuse sections from authentic psychiatric clinical notes, and
3. generate a publicly available vocabulary with substance abuse terms

Our goal is to develop efficient vocabulary generation methods that can be used in larger NLP pipelines for new clinical use-cases, where domain experts with minimal-to-no NLP background can develop tailored solutions for new problems.

1.1 Treatment Management for Acute Anxiety

Patients with depression and anxiety disorders admitted for hospital care commonly receive medications for the management of acute anxiety on an as-needed basis (Curtis and Capp, 2003; Stein-Parbury et al., 2008). These may include benzodiazepines, antihistamines, antipsychotic medications, and others. Although these treatments can reduce a patient's acute distress level, they often have adverse effects. Apart from class-specific side-effects (e.g., oversedation related to benzodiazepines), as-needed anxiolytics may also impair response to psychotherapy and impede long-term recovery (Curran, 1986; Curran and Birch, 1991; Westra et al., 2004; Mystkowski et al., 2003; Otto et al., 2005).

In an effort to better understand the effect of as-needed anxiolytic medications on a patient's ability to manage their anxiety during and after psy-

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 92–101,
Berlin, Germany, August 12, 2016.

chiatric hospitalization, one of the authors (BK) has undertaken a large-scale retrospective study. The study aims to determine whether anxiolytic use correlates with poorer outcomes for psychiatric inpatients being treated for depression and anxiety, such as prolonged hospitalization or increased risk of readmission. This study involves a large cohort (n=about 3000) of patients admitted to several psychiatric hospitals in the same university system. Because the effects of anxiolytic use on the outcomes of interest are likely modulated by a number of other variables, such as her history of substance use disorders, the study requires the coding of almost 30 variables for each patient, all of which must be abstracted from free-text clinical notes.

1.2 Treatment Variables from Clinical Texts

NLP approaches could accelerate the coding process for this data set, while also providing the foundation for future studies with similar aims. Although research in clinical NLP has matured over the last decades, and there are several publicly available clinical text processing pipelines and modules e.g., cTAKES (Savova et al., 2010), MedLee (Friedman et al., 1994), and pyConText (Chapman et al., 2011), adapting and refining these resources to fit the information needs for specific use-cases is not straightforward.

Furthermore, although there have been a few efforts in the NLP community to address mental health-related use-cases, e.g., understanding a patient's suicidal ideations from suicide notes (Pestian et al., 2010) and detecting signals of post-traumatic stress disorder (PTSD), depression, bipolar disorder, and seasonal affective disorder (SAD) from tweets (Coppersmith et al., 2014), NLP for mental health is still in its early stages.

In this study, we focus on substance abuse as it relates to patients suffering from depression and anxiety disorders. As a first step toward encoding substance abuse variables from clinical text, our domain expert (BK) had manually listed a number of terms thought to be relevant to substance abuse. However, these select keywords may not identify all relevant reports due to the variable use of synonyms, abbreviations, acronyms, and misspellings in clinical texts. To assist our domain expert in identifying all relevant patient reports, one initial solution involves automatically extending the initial set of keywords with relevant synonyms and related terms, also known as vocabulary expansion, and marking identified terms for further review. The domain expert (or an NLP module) must then review the report and infer the labels assigned to each substance abuse variable, e.g., from the context of the identified mention "etoh" in the sentence "history of ETOH abuse," assign a label *current*, *past*, *both* or *none* for the variable *Alcohol*.

2 Related Research

The creation of useful domain-specific vocabularies requires a balance between identifying enough terms for adequate coverage (vocabulary expansion) while pruning terms with limited or no utility (vocabulary reduction).

2.1 Vocabulary Expansion

In the biomedicine domain, common vocabulary expansion methods include dictionary-based (e.g., using terminologies and edit-distances), rule-based (e.g., leveraging orthographic/morphological/lexico-syntactic patterns and grammars), machine learning/statistical-based (e.g., applying feature-engineering and transitional states to identify term boundaries), and hybrid approaches (e.g., integrating combinations of the former approaches) (Krauthammer and Nenadic, 2004).

In the clinical domain, vocabulary expansion efforts have included several of these approaches. The Unified Medical Language System, UMLS (Lindberg et al., 1993), has been an influential and important resource for vocabulary development in this domain. Grabar et al. (2009) use the UMLS and other available terminologies to generate synonyms through a compositional analysis along with syntactic dependency information, resulting in high precision (Grabar et al., 2009). Parts of the UMLS have also been enhanced by synonym substitution methods using WordNet (Fellbaum, 1998) and a set of constraints on the number of generated synonyms, resulting in a 10% increase of valid terms related to GI endoscopic examinations in the Minimal Standard Terminology (Huang et al., 2010) Zeng et al. (2012) demonstrated that synonym expansion from the UMLS, topic modeling with Latent Dirichlet Allocation and predicate-based query expansions achieved higher average recalls and average F-measures when compared with the baseline

keyword query for retrieving relevant texts from the United States Veteran Affair's Corporate Data Warehouse (Zeng et al., 2012). Henrikson et al. applied a semi-automatic and language-agnostic method for identifying synonyms of SNOMED CT preferred terms using a distributional similarity technique and a large clinical corpus (Henriksson et al., 2013).

2.2 Vocabulary Reduction

However, many terms from controlled vocabularies like the UMLS and SNOMED CT are not found in biomedical or clinical texts. Hettne et al. (2010) conducted experiments for the building of a medical lexicon using the UMLS Metathesaurus (Hettne et al., 2010). Specifically, they applied term suppression and term rewriting techniques to filter out or discard terms which are considered irrelevant or unlikely to occur in biomedical texts. As a result, a more representative lexicon was produced for medical concept recognition. Wu et al. (2012) conducted a large-scale corpus analysis that leveraged the UMLS Metathesaurus term characteristics to determine which terms generalized across multiple data sources including Mayo Clinic clinical notes and i2b2/VA 2010 NLP Challenge notes, resulting in a set of filtering rules that reduced significantly the size of the original Metathesaurus lexicon (Wu et al., 2012).

Similar to these studies, we aim to assess the utility of terms leveraged from a controlled vocabulary plus a large corpus of notes, and apply various automatic learning approaches to identify relevant terms specifically aimed at an underserved clinical domain. Although some NLP research has addressed the annotation and automatic recognition of variables for substance abuse and its subtopics including *Tobacco*, *Alcohol*, and *Drug Abuse* (Yetisgen et al., 2016; South et al., 2015; Uzuner et al., 2008), to our knowledge, no one has investigated the generation of substance abuse lexicons using our particular term recognition methods for utility in the psychiatric domain.

3 Methods and Materials

In this preliminary study (IRB 68896), we aimed to develop a useful methodology for expanding and reducing domain-specific vocabularies to the most relevant related terms to improve manual patient record review. Our methodology includes three approaches for vocabulary expansion: one ontology-based and two corpus-based (one rule-based using linguistic information and one context-based using neural networks). As a baseline, we used a vocabulary of seed terms defined by a domain expert (BK). Our method also includes an evaluation of characteristics to inform vocabulary reduction: 1) size and content of the generated vocabularies, and 2) coverage and relevance as it relates to authentic data[1]. For the latter, we used a set of psychiatric clinical notes, described below.

3.1 Baseline vocabulary

A set of predefined terms related to substance abuse was used as the baseline vocabulary. These terms were manually generated by a domain expert (psychiatrist, BK) for the purpose of identifying relevant terms from psychiatry notes in relation to specific variables, e.g., term=*opioids*, category=*Opiates*, variable label={*none*, *current*, *past*, *both*}. In total, this substance abuse vocabulary contains 91 terms in 8 categories including e.g., *Alcohol*, *Cocaine* and *Current Smoking Status*.[2] We reference this approach as the *Baseline*.

3.2 Ontology-based vocabulary expansion

To identify relevant synonyms in the UMLS, we searched for each term in the Baseline vocabulary using Knowledge Author (KA) (Scuba et al., 2014). This approach is referenced as *UMLS*.

3.3 Corpus-based vocabulary expansion

Although ontology-based vocabulary expansion approaches can generate many relevant terms, most terms may not be used in practice in clinical texts. A corpus-based approach can be used to identify potentially missed terms and validate the use of ontology-generated synonyms. We used the free-text notes from the entire MIMIC II database (Saeed et al., 2011), which contains clinical documentation for >30000 patients, for two corpus-based vocabulary expansion approaches using: 1) linguistic resources in combination with transformation rules and corpus-based frequency information, and 2) contextual information from a neural network model. Only alphanumeric tokens were

[1]Further details and information about this work, including evaluation script and supplementary material, is available here: `http://toolfinder.chpc.utah.edu/content/vocabulary-expansion-and-reduction-algorithms-vera`.

[2]A subset of a larger vocabulary defined for other variables, e.g. *Education*, *Suicidal Ideation* and *Homelessness*.

used, and all words were converted to lower-case. We included both 1- and 2-token words (uni-and bi-grams) in our methods.

Linguistic and rule-based approach generates lexical variants by querying each seed term (e.g., "alcohol abuse") in WordNet after which four steps are applied on each generated WordNet synonym: 1) term reordering ("abuse alcohol"), 2) inflection generation ("alcohol abused"), 3) abbreviation generation ("aa") and 4) typographic error generation ("alchol abuse"). Each generated term variant was then checked against the MIMIC II corpus and candidate terms occurring >15 times were kept (Conway and Chapman, 2012). We reference this approach as *WNLing*.

Neural network approach leverages a word2vec (Mikolov et al., 2013) neural network bigram model for the generation of context-based related terms. We built a model using a window parameter of 5, discarded words occurring < 15 times, and set the vector dimensionality to 400. Each term in the baseline vocabulary was then queried to find the most similar uni- or bigrams with a similarity score $>= 0.5$.[3] This approach is called *word2vec*.

3.4 Evaluation data set

We randomly sampled 100 psychiatric clinical notes (from a total of approx. 2500) from the University Hospital, University of Utah, Salt Lake City, collected for the purpose of extracting information related to *as-needed anxiolytic use*. From each note, sections more likely to contain information about substance abuse (e.g., *PSYCHIATRIC HISTORY AND PHYSICAL* and *PSYCHIATRIC H&P*) were extracted for matching terms from each vocabulary.

3.5 Evaluation

We performed a quantitative evaluation from two perspectives: 1) vocabulary size and content, to understand characteristics of the generated vocabularies, and 2) vocabulary coverage and relevance, to understand their applicability on authentic data.

3.5.1 Vocabulary size and content

Each new vocabulary was generated from the list of terms in the Baseline vocabulary[4]. We calculated the number of terms in each generated vocabulary, the number of added terms as compared to the other vocabularies, as well as the total number of all terms (set union) and the total number of shared terms (set intersection) between the generated resources. Note that the vocabularies may contain unigrams that are parts of larger n-grams (multi-word tokens e.g., "alcoholics" as a part of "alcoholics anonymous"). Each unique term was counted separately.

3.5.2 Vocabulary coverage and relevance

Each term in each generated resource was matched against the evaluation data to calculate number of terms found and frequencies of occurrence. A simple string matching procedure was employed in each substance abuse section using regular expressions, where a match was counted if a term[5] was found between a word boundary (*"\b"*).

As this evaluation data set is not manually annotated for substance abuse-related terms, we instead calculated *approximations* for both precision (positive predictive value) and recall (sensitivity) by comparing the terms generated from each approach to terms generated by all four approaches.

To calculate these versions of precision and recall for each approach, *relevant and correct* terms (true positives, TP) were defined as the set union of the pairwise intersection sets between all four approaches, i.e. all terms that were found by at least two approaches. *Missed* terms (false negatives, FN) were defined as the terms *not* generated by a specific approach but generated by one (or more) combination of other approaches. *Spurious* terms (false positives, FP) were defined as the number of terms found by a specific approach, but not any other approaches.

Precision and *recall* were then calculated from these results for each approach (*precision=TP/(TP+FP)* and *recall=TP/(TP+FN)*), respectively. Note that this approximation ought to be analyzed with caution - it only gives results in relation to the terms that the vocabularies generate (there is no knowledge about potentially relevant terms outside of these vocabularies). Since the vocabulary approaches are rather different, we believe that this evaluation does give a hint towards what could be expected to at least be relevant terms, and illustrates the relationship between

[3] We used the gensim package (Řehůřek and Sojka, 2010) to build this model.

[4] Note that some terms in the Baseline vocabulary were not found in the generated models, Table 7 in Supplement: `http://toolfinder.chpc.utah.edu/sites/default/files/psychiatry_substance_use_related_terms_supplement.pdf`.

[5] excluding English stopwords from the nltk (http://www.nltk.org/) package.

the employed approaches. This evaluation also permits us to learn common terms learned by multiple approaches and contemplate which combinations should be presented back to the domain expert for expanding the initial query.

4 Results

We report characteristics of the vocabularies in terms of size and content, and we report on vocabulary coverage and relevance when applied to the evaluation data.

Vocabulary	size
Baseline	91
UMLS	863
WNLing	1253
word2vec	1758

Table 1: Size of each vocabulary (number of unique terms).

4.1 Vocabulary size and content

The number of terms in each vocabulary (Baseline, UMLS, WNLing and word2vec) is reported in Table 1. Excluding the Baseline, the word2vec model generated the most terms (n=1758), while UMLS generated the fewest (n=863). In total, 3661 unique terms were generated (Baseline ∪ UMLS ∪ WNLing ∪ word2vec).

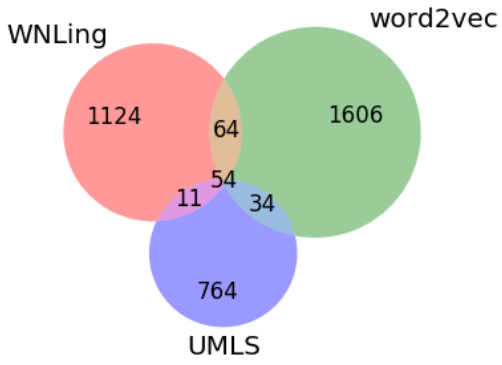

Figure 1: Venn diagram: number of terms in the generated vocabularies from the three approaches: UMLS, WordNet with linguistic heuristics (WNling), and word2vec.

One term from the Baseline vocabulary was not found by any of the other approaches (*substance use history*). For the corpus-based approaches, there were also some terms in the Baseline vocabulary that were not present in the models generated from the MIMIC II corpus[6].

Forty-three terms were shared between all four vocabularies (Baseline ∩ UMLS ∩ WNLing ∩ word2vec)[7]. and eleven additional terms (a total of 54) were shared between the three approaches (UMLS ∩ WNLing ∩ word2vec): *addictions*, *alcoholic beverage*, *alcoholic drink*, *amphetamines*, *beer*, *benzodiazepines*, *drug abuse*, *ethanol*, *ethyl alcohol*, *glass*, and *substances*.

Figure 1 shows a Venn diagram with the results from the three vocabulary expansion approaches (UMLS, WNLing, word2vec). In total, 163 terms were shared between at least two approaches (182 in total when including the Baseline vocabulary). Among these 163 terms, added terms as compared to the Baseline vocabulary include misspelling variants (*morpine*, *coccaine*), inflections (*smokers*, *addictions*) in addition to new, potentially relevant terms such as *narcotic*, *etoh*, *codeine*. The proportion of shared terms for each pairwise vocabulary combination roughly reflects the sizes of the vocabularies, e.g. word2vec ∩ WNLing (n=64) > UMLS ∩ WNLing (n=11).

4.2 Vocabulary coverage

Vocabulary	u	tot	min	max	avg
Baseline	37	416	1	11	4
UMLS	49	536	2	11	5
WNLing	85	828	2	18	8
word2vec	104	786	1	21	7

Table 2: Number of terms found in the evaluation data. u = number of unique terms found irrespective of frequency, tot = total number of term occurrences found, min, max, avg = minimum, maximum and average number of terms per section.

The number of matched terms (unique and total) in the 100 random substance abuse sections from each vocabulary is shown in Table 2. The sections contain in total 4036 words[8] (min=4, max=192, avg=40). We observed an average of 4 (Baseline) to 8 (WNLing) substance abuse terms (min: 1;

[6]Table 7 in Supplement.
[7]Table 1 in Supplement.
[8]Counted using a simple whitespace tokenizer

max: 21) in each substance abuse section using the different vocabularies, Table 2.

The proportion of observed unique terms found in the substance abuse sections varied from about 5-7% for UMLS, WNLing, and word2vec compared to about 41% for the Baseline. As the size of the vocabularies increased, so did the total number of term occurrences (about 7.5 to 11 fold).

A comparison of the coverage between the generated vocabularies is depicted in the Venn diagram in Figure 2. Overall, the number of matched terms is higher for the larger vocabularies (WNLing, word2vec) and the proportion of shared terms is also higher. Twenty-eight terms are shared between all three new vocabularies, and 52 terms are shared between at least two vocabularies (16+28+2+6).

To evaluate the approximated precision and recall, we use the union of each pairwise intersection of *all* vocabularies (including the Baseline), which resulted in 57 unique terms[9].

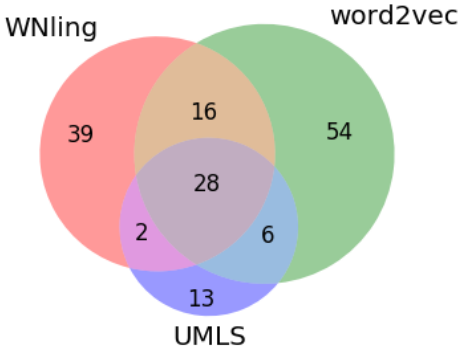

Figure 2: Venn diagram: number of unique terms from the generated vocabularies found in the evaluation data: UMLS, WordNet with linguistic heuristics (WNling), and word2vec.

As expected, the Baseline approach resulted in the highest approximated macro-and micro-precision (0.97/0.998), Table 3. In contrast, the vocabulary-based approaches resulted in the highest macro-and micro-recall (0.91/0.98: word2vec). The fact that the micro-results are higher for the two corpus-based approaches indicates that these two approaches generate term variants that are also more frequent in the evaluation data set.

4.3 Vocabulary relevance

To analyze relevance, the 20 most frequent terms found by each approach is presented in Table 4, along with information about which vocabulary the term was found in. Nine of these terms were found in all vocabularies. Two terms were not clearly relevant to substance abuse (*last*, *years*). The three most frequent terms, that were found in all 100 substance abuse sections, are uni- and bi-gram variants of the same term (*substance use*, *substance* and *use*).

Vocabulary	Macro		Micro	
	P	R	P	R
Baseline	0.97	0.63	0.998	0.67
UMLS	0.77	0.67	0.78	0.68
WNLing	0.55	0.82	0.69	0.93
word2vec	0.5	0.91	0.77	0.98

Table 3: Results: approximated precision and recall, macro (per unique term) and micro (per term occurrence). The number of relevant (and correct) terms is defined as the set union of all pairwise intersections.

The Baseline vocabulary included e.g. the term *packs* but not its singular inflection *pack*, which turned out to be more frequent (*pack* freq=18 as opposed to *packs* freq=6[10]. and found by the two corpus-based approaches WNLing and word2vec. Each approach also resulted in a number of potentially relevant terms that were not found in any of the other approaches, e.g. *amphetamine abuse* (UMLS), *withdrawal* (WNLing), *demerol* (word2vec)[11].

5 Discussion and Conclusion

We present a simple and systematic approach for automated vocabulary generation (expansion and reduction) in the domain of *substance abuse*, applied and evaluated on a set of substance abuse sections from authentic psychiatric notes. Three vocabularies were generated from a set of seed terms using publicly available resources (ontologies, software, and corpora) and combined to: 1) generate a substance abuse vocabulary of highly relevant terms and 2) characterize and analyze

[9]Table 2 in Supplement.

[10]Table 2 in Supplement.

[11]Tables 3–6 in Supplement.

term	Baseline	UMLS	WNLing	word2vec	freq
substance use		x			100
substance	x	x	x	x	100
use			x		100
alcohol	x	x	x	x	62
history			x		41
drug			x	x	40
abuse	x		x	x	40
tobacco	x	x	x	x	35
marijuana	x	x	x	x	32
smokes	x	x	x	x	27
drug use		x		x	26
drug abuse		x	x	x	23
illicit drug				x	22
last			x		21
cocaine	x	x	x	x	21
cigarettes	x	x	x	x	19
pack			x	x	18
years				x	15
heroin				x	15
smoking	x	x	x	x	14

Table 4: 20 Most frequent terms from the union set of all vocabularies that were found in the evaluation data. Presence of term in each respective vocabulary is marked with "x". Note that unigrams could be a substring of an n-gram in each vocabulary (e.g. *substance* and *substance use* in the UMLS vocabulary).

coverage and relevance in an authentic psychiatric dataset.

Through our definition of an approximated precision and recall, we observed that the baseline and ontology-based approaches resulted in the highest approximated precisions, suggesting these methods are useful for identifying the most relevant related terms. This finding is not surprising because the list was vetted by a domain expert and core to the set of terms for all four approaches. In contrast, the vocabulary-based approaches resulted in the highest recalls suggesting these methods are useful for identifying potentially new related terms.

The denominator for calculating these results was based solely on a combination of the four generated vocabularies, which only illustrates relations between approaches. Interestingly, the ontology-based approach (UMLS) resulted in moderate performance for both precision and recall. We hypothesize that this result occurs because although the UMLS provides a notable number of unique terms, these terms do not frequently occur in clinical text due to term characteristics (e.g., inclusion of semantic type and special characters) and concept granularity (use of chemical nomenclature for specific drugs). In future work, we will apply methods to filter based on these characteristics similar to (Hettne et al., 2010; Wu et al., 2012; Demner-Fushman et al., 2010) to address these and other challenges with knowledge authoring leveraging noisy resources.

The baseline vocabulary was biased to more specific terms of substance abuse usage including terms for substances (*alcohol*, *marijuana*, *tobacco*, *cocaine*, and *cigarettes*). Both the ontology- and corpus-based approaches identified more general terms for substance abuse and drug usage as well as terms related to linguistic/semantic attribute information (e.g. *drinking heavily*, *rarely drinks*, *quit smoking*). In future work, we will develop methods for learning these patterns to infer these attribute information. These methods will then be placed into a larger infrastructure called *the Information Extraction-Visualization pipeline (IE-Viz)* to aid domain experts with no-to-minimal NLP experience in developing NLP systems for domain-specific use

cases.

A preliminary manual analysis of the resulting list of relevant terms (true positives) revealed that a clear majority of the resulting terms were related to substance abuse - only one term was obviously problematic (*years*). The false positives, on the other hand, were in many cases actually relevant and correct terms (e.g. *ecstasy* from the word2vec model), although the WNLing model also produced a number of irrelevant terms (e.g. *charges*). To assess our approximated coverage and relevance metrics, we will conduct a manual assessment of the performance of this approach with respect to true coverage, i.e. analyze which terms were missed, as well as correctness for found terms to determine how well our approximated precision and recall corresponds to the actual precision and recall of terms from this evaluation data set. Moreover, we will assess the relation between terms and categories.

We aim to extend our vocabulary expansion and reduction methods. Most importantly, we have only performed one iteration, using domain expert curated terms, to create the final list of terms. This list could be extended by performing a number of iterations on the resulting list, thereby generating a richer and more comprehensive set of terms. Moreover, we plan to utilize additional publicly available resources, e.g. relevant Wikipedia pages. Once these methods have been integrated into our NLP pipeline, we will extend our experiments to the other psychiatric variables from our data set, e.g., social risk factors of *Homelessness*, *Education level*, *Abuse as a Child*, *Suicide attempts/Self Harm*, and new clinical use-cases such as the detection of bleeding events associated with anticoagulant medication usage by patients with high-risk of stroke.

5.1 Limitations

Our preliminary study evaluation has several limitations including an evaluation using a small data set and calculation of term matches without consideration of term overlap (unigrams/multi-word token counts). We aim to extend our evaluation data set and calculate the effect of term matching criteria in follow up work.

Some of our vocabulary expansion methods have limitations and might be improved. Specifically, word2vec and similar approaches generate related terms that could be a relatedness of semantic types other than synonyms (antonyms, hyponyms, etc.), which is well-known. However, we believe co-occurrence of these terms may correlate with variable terms and perhaps subsequent labels, e.g., alcohol occurring with smoking, which may help with extraction efforts in our NLP pipeline downstream. WNLing abbreviation methods can generate many false positives. Although we reduced some false positives with a stopword check, we could leverage medical acronym and abbreviation dictionaries such as the *Medilexicon*[12] and the *STANDS4 network*[13] to further reduce false positives. Moreover, we believe that combining these types of approaches can be a useful way of limiting the impact of each method's disadvantages.

Finally, our thresholds were chosen rather arbitrarily; therefore, we will experiment with determining the effect of similarity scores and word count thresholds, as well as the use of larger *n*-grams.

5.2 Contribution

To our knowledge, this study is the first systematic study of terms related to substance abuse generated from publicly available resources and the combinations of these approaches, and then evaluated on authentic psychiatric notes. The generated vocabularies can be used to automate parts of the variable encoding process for the ongoing study on treatment management of hospital admitted patients with depression and anxiety disorders, as well as other clinical use-cases where substance abuse information is of importance. This work represents a first step in a larger framework to empower domain experts, in this case psychiatrists, to develop queries and apply NLP methods to identify and extract substance abuse and other variables from large clinical data sets to support mental health research.

Acknowledgments

We would like to thank the anonymous reviewers for valuable comments. This work is partly funded by the Department of Veteran Affairs (CRE 12-312), the National Library of Medicine (R00LM011393), the Patient-Centered Outcomes Research Initiatives (CDRN-1306-04912), the Swedish Research Council (2015-00359), and the

[12] http://www.medilexicon.com/medicalabbreviations.php
[13] http://www.abbreviations.com/about.php

Marie Skłodowska Curie Actions, Cofund, Project INCA 600398.

References

Brian E. Chapman, Sean Lee, Hyunseok Peter Kang, and Wendy Webber Chapman. 2011. Document-level classification of CT pulmonary angiography reports based on an extension of the context algorithm. *Journal of Biomedical Informatics*, 44(5):728–737.

Mike Conway and Wendy W. Chapman. 2012. Discovering Lexical Instantiations of Clinical Concepts using Web Services, WordNet and Corpus Resources. In *AMIA 2012 Proceedings*, page 1604, Chicago, USA, November. American Medical Informatics Association.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Association for Computational Linguistics Workshop of Computational Linguistics and Clinical Psychology*.

Helen V. Curran and Brian Birch. 1991. Differentiating the sedative, psychomotor and amnesic effects of benzodiazepines: a study with midazolam and the benzodiazepine antagonist, flumazenil. *Psychopharmacology (Berl)*, 103(4):519–23.

Helen V. Curran. 1986. Tranquillising memories: a review of the effects of benzodiazepines on human memory. *Biol Psychol*, 23(2):179–213, Oct.

Janette Curtis and Kim Capp. 2003. Administration of 'as needed' psychotropic medication: a retrospective study. *Int J Ment Health Nurs*, 12(3):229–34, Sep.

Dina Demner-Fushman, James G Mork, Sonya E Shooshan, and Alan R Aronson. 2010. UMLS content views appropriate for NLP processing of the biomedical literature vs. clinical text. *J Biomed Inform*, 43(4):587–94, Aug.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech and Communication. Mit Press.

Carol Friedman, Philip O. Alderson, John H. Austin, James J. Cimino, and Stephen B. Johnson. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association : JAMIA*, 1(2):161–174, March.

N. Grabar, PC. Varoutas, P. Rizand, A. Livartowski, and T. Hamon. 2009. Automatic acquisition of synonym resources and assessment of their impact on the enhanced search in EHRs. *Methods Inf Med.*, 48(2).

Aron Henriksson, Mike Conway, Martin Duneld, and Wendy Webber Chapman. 2013. Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records. In *AMIA 2013, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 16-20, 2013*.

Kristina M Hettne, Erik M van Mulligen, Martijn J Schuemie, Bob Ja Schijvenaars, and Jan A Kors. 2010. Rewriting and suppressing UMLS terms for improved biomedical term identification. *J Biomed Semantics*, 1(1):5.

Kuo-Chuan Huang, James Geller, Michael Halper, Yehoshua Perl, and Junchuan Xua. 2010. Using WordNet Synonym Substitution to Enhance UMLS Source Integration. *Artif Intell Med.*, 46(2).

Michael Krauthammer and Goran Nenadic. 2004. Term identification in the biomedical literature. *J Biomed Inform*, 37(6):512–26, Dec.

DA. Lindberg, BL. Humphreys, and McCray AT. 1993. The Unified Medical Language System. *Methods Inf Med.*, 32(4):281–91.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*. NIPS.

Jayson L Mystkowski, Susan Mineka, Laura L Vernon, and Richard E Zinbarg. 2003. Changes in caffeine states enhance return of fear in spider phobia. *J Consult Clin Psychol*, 71(2):243–50, Apr.

Michael W Otto, Steven E Bruce, and Thilo Deckersbach. 2005. Benzodiazepine use, cognitive impairment, and cognitive-behavioral therapy for anxiety disorders: issues in the treatment of a patient in need. *J Clin Psychiatry*, 66 Suppl 2:34–8.

John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. 2010. Suicide note classification using natural language processing: A content analysis. *Biomedical Informatics Insights*, (3):1928.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. 2011. Multiparameter intelligent monitoring in intensive care ii: a public-access intensive care unit database. *Crit Care Med*, 39(5):952–60, May.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping. Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

William Scuba, Melissa Tharp, Yang Tseytlin Eugene, Liu, Frank A. Drews, and Wendy Chapman. 2014. Knowledge Author: Creating Domain Content for NLP Information Extraction. In *6th International Symposium on Semantic Mining in Biomedicine (SMBM)*.

Brett R. South, Danielle Mowery, Melissa Tharp, Margorie Carter, Adi Gundlapalli, Marzieh Vali, Mike Conway, Salomeh Keyhani, and Wendy W. Chapman. 2015. Extracting social history and functional status from veteran affairs clinical documents. In *AMIA Joint Summits on Translational Science*.

Jane Stein-Parbury, Kim Reid, Narelle Smith, Diane Mouhanna, and Fiona Lamont. 2008. Use of pro re nata medications in acute inpatient care. *Aust N Z J Psychiatry*, 42(4):283–92, Apr.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac S. Kohane. 2008. Viewpoint paper: Identifying patient smoking status from medical discharge records. *Journal of American Medical Informatics Association*, 15(1):14–24.

Henny A. Westra, Sherry H. Stewart, Michael Teehan, Karen Johl, David J. A. Dozois, and Todd Hill. 2004. Benzodiazepine Use Associated with Decreased Memory for Psychoeducation Material in Cognitive Behavioral Therapy for Panic Disorder. *Cognitive Therapy and Research*, 28(2):193–208, April.

Stephen Tze-Inn Wu, Hongfang Liu, Dingcheng Li, Cui Tao, Mark A. Musen, Christopher G. Chute, and Nigam H. Shah. 2012. Unified medical language system term occurrences in clinical notes: a large-scale corpus analysis. *Journal of Americal Medical Informatics Association*, 19(e1).

Meliha Yetisgen, Elena Pellicer, David R. Crosslin, and Lucy Vanderwende. 2016. Automatic identification of lifestyle and environmental factors from social history in clinical text. In *AMIA 2016 Joint Summits on Translational Science*.

Qing T. Zeng, Doug Redd, Thomas Rindflesch, and Jonathan Nebeker. 2012. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. volume 2012, pages 1050–1059. American Medical Informatics Association.