Towards Creating an Iso-semantic Lexicon Model using Computational Semantics and Sublanguage Analysis Within Clinical Subdomains for Medical Language Processing

B. S. Begum Durgahee[1,3] and Adi Gundlapalli[1,2,3]
[1]*Department of Biomedical Informatics, University of Utah, Salt Lake City, U.S.A.*
[2]*Department of Internal Medicine, University of Utah, Salt Lake City, U.S.A.*
[3]*VA Salt Lake Health Care System, Salt Lake City, U.S.A.*

1 STAGE OF THE RESEARCH

The current research proposal is at the initial stage. Some preliminary investigations have already been completed and currently working on the first aim of the proposal.

2 OUTLINE OF OBJECTIVES

The main objective of this research proposal is to leverage semantic lexicons for improving Natural Language Processing (NLP) performance and interoperability. The integration of sublanguage specific patterns and relations in the design of learning-based systems with existing ontologies are promises for facilitating clinical research informatics.

Hypothesis. Semantic lexicons can be leveraged to facilitate content interoperability for information extraction from clinical texts.

Specific Aims

1. Identifying lexico-semantic relations and patterns from clinical texts from a large health care system by investigating contextual linguistic knowledge from the clinical domain.
2. Building an adaptive Information Extraction application for creating a common semantic based lexicon model by using features identified from aim 1. and other sources, unsupervised learning methods and existing ontologies and terminologies.
3. Building of ontology-based semantic lexicon, using the output from aim 2 using Semantic Web technologies and lexical ontologies for facilitating data integration.

3 RESEARCH PROBLEM

Currently homelessness is a serious issue in the United States. Homelessness is associated to socioeconomic factors such as cost of living, unemployment, and poverty, in addition to individual factors, like mental illness, behavioral factors and family issues. However, homelessness is more frequent among the Veterans due to exposure to military deployment. The Veterans are more prone to be patients suffering from physical or mental disabilities as a result of injuries from military missions and eventually this may be associated towards loss of jobs and substance abuse. Therefore it is essential to identify these patients at an early stage to better manage and prevent homelessness. Unfortunately, identifying evidence and risk factors associated to homelessness are not straight forward since structured medical data, namely International Classification of Diseases (ICD) codes are not always complete nor representative of patient complaints and their clinical state (Shivade et al., 2014; Birman-Deych et al., 2005). Therefore NLP systems are required to transform clinical narratives (unstructured data) into a suitable form for informatics tasks (structured data) (Friedman et al., 2013). However, NLP systems are faced with challenges like misspellings, redundancy, ungrammatical texts, abbreviations, dialectal short phrases, ambiguity, boilerplate and template from the clinical narratives (Meystre et al., 2008).

Prior work has demonstrated the use of semantic lexicons in NLP systems to benefit the analysis of medical narratives (Johnson, 1999; Liu et al., 2012a; Jonnalagadda et al., 2013). A semantic lexicon associates clinical words and phrases to appropriate concepts for medical language processing. However, the lexicon building process has mostly been performed in silos and according to user needs or specifications (Shivade et al., 2014). Besides being inherently het-

Durgahee B. and Gundlapalli A..
Towards Creating an Iso-semantic Lexicon Model using Computational Semantics and Sublanguage Analysis Within Clinical Subdomains for Medical Language Processing.

erogeneous, structured and labor-intensive, there is no standardized way of building the lexicons in the clinical domain. Researchers have looked at ways to automate medical lexicon construction using existing standard(Johnson, 1999; Jonnalagadda et al., 2013). Controlled vocabularies from the Unified Medical Language System (UMLS) Metathesaurus have been used to construct semantics lexicons, but are effective for NLP if a lexeme has one semantic type. Johnson (Johnson, 1999) had to apply semantic preference rules to alleviate semantic ambiguity among the medical terms mapped between the Specialist Lexicon and UMLS Metathesaurus. The UMLS sources and individual standard terminologies such as SNOMED-CT, CPT and LOINC, may have limited coverage for observed usage of terms and variants of clinical concept(Wu et al., 2012; MacLean and Heer, 2013; Friedlin and Overhage, 2011)

The use of biomedical ontologies to tag entities in the domain is not sufficient (MacLean and Heer, 2013); polysemic words in sublanguages may belong to different semantic classes. Many of the medical terms undergo shifts of meaning depending on surrounding contextual words. Moreover, nuances, ambiguity and presupposition may exist in clinical notes that are associated to the facility, specialty and provider respectively. The semantic lexicons tend to be restricted to the vocabulary being used by an environment and eventually result in different research groups implementing their own NLP algorithms. As a result, the performance of NLP or IE Systems then varies across institutions, providers and sources of data and cannot be easily compared. The challenge is to establish semantic interoperability of clinical lexical resources from different sites into a standard form for reuse by NLP applications. There is a need to have stronger connection between ontological concepts and text level information to facilitate sense alignment between clinical resources.

Recently, NLP has been efficient in extracting entities such as persons, procedures, drugs, diseases and genes (Abacha and Zweigenbaum, 2011; Huang et al., 2012; Jonnalagadda et al., 2013). However, little research has focused on identification of information structures, predicates between identified entities and events underlying the clinical text. Therefore the proposed research study aims to leverage semantic lexicons that are accurate, augmented with linguistic properties and that can be shared for information extraction from clinical notes.Ultimately, we hope that the study will improve the efficiency of IE performance to retrieve reliable clinical information that can help in personalizing care to prevent homelessness.

4 STATE OF THE ART

Despite the increased use of Electronic Health Records (EHR) and attempts to bring structure to provider's written notes using templates and drop-down menus, large amounts of clinical information are represented in narrative free-text form. Efforts to understand and parse unstructured clinical notes have been an ongoing research problem for Natural Language Processing and Information Extraction (IE) researchers for several decades (Meystre et al., 2008). However, little research has been done towards bridging the semantic interoperability gap in medical language processing.

Semantic ambiguity arises when a clinical term has multiple semantic classes and when the sense of the term varies according to context and usage within the text. Therefore, NLP applications require a semantic lexicon that has normalized the mapping of similar terms to the same concept for semantic analysis of the text. Being an integral part of the NLP process, the semantic lexicon plays a crucial role in defining the type of informatics task that NLP need to perform. For example, to identify patient cohorts at risk for homelessness requires a semantic lexicon mostly containing term-concepts mapping associated with homelessness and related psycho-social factors, while patients at risks for diabetes will have term-concepts mapping associated to family history, gender and weight, among others.

Typically, in the clinical domain the lexicons for natural language processing are mostly constructed by hand or from existing medical vocabularies. Insufficient linguistic representations and incompleteness of terminologies in the clinical domain are a few of the challenges in building semantic lexicons (Liu et al., 2012a; MacLean and Heer, 2013; Verspoor, 2005). Although, there is an increase in the number of vocabularies and NLP applications in the biomedical domain, it is contended that the semantic interoperability problem in medical language processing has been augmented due to isolation and lack of integration of biomedical resources. Recently constructed semantic lexicons for NLP systems are restricted to specific phenotyping criteria, thus impeding reuse and maintainability.

4.1 Sublanguage and Unsupervised Information Extraction

Lexicons have been recognized as a critical component of NLP systems (Guthrie et al., 1996). Selectional restrictions are not sufficient to represent all the semantic information for NLP systems. Humans are

linguistically creative and paraphrase concepts in various ways. Therefore, words are associated to each other in many different styles, such that this information does not exactly match predefined patterns in standardized ontologies or terminologies. The challenge lies in deducing a priori those inherent inequalities of the likelihood of certain pattern of words occurring in clinical notes.

The presence of sublanguages within the clinical domain has been recognized for more than a decade (Friedman et al., 2002; Sager et al., 1995). A sublanguage exists within a text when a limited group of people includes its own vocabulary, syntactic and semantic characteristics. Previous studies have used sublanguage-approaches for clinical information extraction, however these systems adapt poorly to different domains (Meystre et al., 2008). Most of these systems are rule-based and restricted to a few relations from the plethora of undiscovered information structures in clinical notes. Moreover, the concept extraction process of these systems is data driven by specific tasks and vocabulary used at specific sites.

One of the fundamental secondary use of clinical notes is for phenotyping purposes. Identifying patients for phenotyping is dependent on specific tasks and hence NLP systems are customized according to handcrafted lexicons for these tasks. This tends to be expensive and is an impediment for interoperability. Recently, there has been some progress towards using structural linguistics with supervised machine learning for clinical information extraction (Jonnalagadda et al., 2013; Uzuner et al., 2011). Supervised learning methods using clinical data have been slow due to limited access to annotated clinical data for training. Uzuner et al., showed that supervised methods with some rules accurately identified relations between medical concepts (Uzuner et al., 2011). However when unannotated training data are inadequate, ensemble of classifiers, information from unlabeled data and other sources may be required. Similarly, Jonnalagadda et al. used a supervised method with distributional information to build a lexicon of treatments, tests and medical problems from clinical notes and biomedical literature (Jonnalagadda et al., 2013). The authors demonstrated that their technique could supplement or replace manually created lexicons.

The scarcity of training data can be overcome by unsupervised IE systems. Adaptive IE and Open domain IE are two promising directions for unsupervised learning (Grishman, 2001). Open domain IE does not depend on predefined vocabulary while adaptive IE can be achieved through limited parsing and learned patterns from word sequences. By learning the surrounding contexts, the IE system can be adapted to a particular domain. Grishman (Grishman, 2001) supports the premise that the integration of structured linguistics have the potential for developing adaptive IE. He also raises concerns regarding the need for innovative distributional analyses for building information structures above the kernel level. Previous studies have confirmed that the language used in clinical notes differs from general English or biomedical text (Coden et al., 2005; Friedman et al., 2002; Meystre et al., 2008). The sublanguage difference in clinical narratives has been shown to be not only across non-medical areas, but within clinical sub-domains as well. Patterson and Hurdle (Patterson and Hurdle, 2011) demonstrated that specific sublanguages were distinguished depending on the extent of the clinical sub-domain. They proved that note types cluster according to the extent of the clinical subdomain. Similar to (Wu et al., 2012), this study suggests that term frequencies and semantics information may depend on the institution and source of the clinical notes. Therefore, domain adaptation may be considered a key component for the development of efficient clinical information extraction tools.

4.2 Computational Semantics in the Biomedical Domain

The growing number of semantic and pragmatic characteristics found in clinical notes forms the foundation of Computational Semantics for medical language processing. The clinical note consists of real world physician-patient interaction descriptions. In addition to semantics, domain knowledge, contextual information and inferences are among other factors contributing to the clinical narrative. Stephen Wu (Wu, 2013), endorses the view that the relatively new sub-discipline of NLP, Computational Semantics, has a crucial role in medical informatics. In his paper, the author urges the development of new Computational Semantics techniques for clinical text processing, since the ultimate goal is to capture the semantic meaning associated to medically-related entities, such as signs or symptom or disease and problems. In general, Computational Semantics is a technique that automatically builds semantic representations from language expressions by integrating formal semantics, computational linguistics and reasoning.

Some recent work has shown great promise in using computational semantics on clinical notes for patient outcome predictions, name entity recognition and vocabulary construction (Howes et al., 2013; Jonnalagadda et al., 2013; Kate, 2013; Sohn et al., 2013; Zweigenbaum et al., 2013). Howes, Purver and McCabe (Howes et al., 2013) used the Latent Dirich-

let Allocation (LDA) probabilistic modelling to investigate whether topic modeling can predict patient outcomes, such as symptoms and/or therapy. The authors deduced that LDA predicted therapeutic relationship evaluations in a more efficient manner as compared to symptoms. This is supported by the fact that automatic modeling techniques, such as LDA, are more suitable for identifying content style and structure rather than the concepts. However, this shows the potential of applying unsupervised learning methods to cluster similar communication style or structure within a corpus. Similarly, Jonnalagadda et al. and Zweigenbaum et al. explored computational semantics methods for named entity recognition (NER) (Jonnalagadda et al., 2013; Zweigenbaum et al., 2013). The former group investigated distributional semantic methods for extracting treatment, tests and medical problem entities from clinical notes and the biomedical literature. Their focus was to limit the use of high-cost resources, such as annotated data. The distributed semantic features, n-nearest word, support vector machine (SVM), and term clustering, were used to automatically build domain independent lexicons for NER. Although, the authors maintain that manually constructed lexicons should be used when available, distributional methods can alleviate the handcrafted and tedious effort when building lexicons from scratch. The clinical NER had greater impact when distributional semantic features were used rather than the manual lexicon. The F-score was increased by 2.0%, showing potential in automatic clinical lexicon construction with minimal human effort from corpora. However, they were concerned regarding the performance of their method across different clinical corpora from different sources or providers. Similarly Zweigenbaum et al. (Zweigenbaum et al., 2013) investigated entity recognition from clinical notes, but focusing towards combining expert knowledge with data-driven methods. They showed that the F-measure for clinical NER increased by just using the output of the expert-based system as features to the Conditional Random Field (CRF) classifier. However, they argue that their methodology may not be generalizable across domains and sources.

The interesting observation from the above studies is that distributional information was explored as features with their respective machine learning method. Although, the increase in performance from these studies is not substantial from their baselines, it does show that distributional information has the potential to improve NLP understanding. A different approach was taken by Kate et al. to identify SNOMED-CT relations to eventually convert clinical phrases from text to SNOMED-CT expressions (Kate, 2013). The authors showed significant improvement in identifying relations from clinical text using SVM and a novel kernel approach. They show that their method has the potential to identify new clinical SNOMED-CT expressions. Moreover, the formalization of SNOMED-CT into description logic (DL) creates opportunities for automated reasoning, which is a fundamental part of Computational Semantics. To the best of my knowledge, there is no work, which has attempted to explore inferencing using DL in clinical text.

Distributional methods seem to be suitable for automatic generation of semantic lexicons. However, the performance of distributional methods on its own does not capture all relations, more specifically those derived through inferences. Therefore, combining different techniques is an alternative research direction for improving performance of IE. Methodologies such as graph based models and functional similarity measures are more focused towards discovering of new knowledge through inferences. In the Biomedical Literature, gene and protein names were extracted by combining different approaches including pattern-based, distributional and graph-theory (Cohen et al., 2005). Network structure was created based on the fact that similar gene or protein names will have denser networks than unrelated terms. Recently, graph based model, such as random walk model, was applied to determine whether two words or phrases are related by using features from Wikipedia (Zhang et al., 2010). Six Wikipedia features, weighted according to a semantic relatedness measure, were used to disambiguate name entities from news stories. The authors evaluated their novel approach and deduced that Wikipedia performed better than using WordNet-based approaches. Another study which was performed by Pedersen and colleagues (Pedersen et al., 2007), demonstrated that existing domain-independent measures such as path-based and vector based methods could be applied in the biomedical domain to find similar concepts. However, they argue that manually constructed ontologies on their own may not be sufficient to cover all semantic relationships in the clinical corpora. This is due to the fact that clinical text may contain presupposition and a substantial amount of contextual information. Therefore, they conclude that path-based measures integrated with information content statistics from corpora can help in capturing contextual information as well as be a promising methodology for domain adaptation.

Different approaches have been discussed above, and the clinical corpora provide a number of opportunities for Information Extraction and knowledge dis-

covery. However, an important requirement these days is that IE systems should be designed such that they can easily adapt to new domains to automatically extract relevant relations or events. Nowadays, the number of ontologies available for the Biomedical Domain are increasing and growing in size. One such example is SNOMED-CT, which is the largest source of medical concepts that is used worldwide. However, these existing sources of medical terminology has not been exploited enough for knowledge discovery. More specifically, combining learning methods with ontologies is an interesting area of research, where text usage information from corpora could assist with predicting new terms or conceptual lexical variants in the clinical domain. The time is ripe to propose and develop novel approaches that take advantage of biomedical ontologies, description logics and corpus based statistical measures for discovering new terms in the clinical text.

4.3 Semantic Web, Clinical Ontologies and Lexicons

The Semantic Web enables sharing of machine-usable content beyond the boundaries of applications and websites. The term, 'Semantic Web', indicates that semantics is the center of this technology, and more specifically towards facilitating semantic interoperability for information exchange. This can be achieved by formal ontologies that help in structuring data in a way that machine can understand. In the clinical domain, the number of vocabularies keeps increasing. ICD, MeSH, SNOMED, and NCI Metathesaurus are the most common clinical ontologies and terminologies. Although these vocabularies have similar sub-language and structural representation, the granularity of the representations differ across the databases. Moreover, the different clinical terminologies and ontologies are constructed in silos, hence creating integration complexities. ICD-9 consists of a quarter of the amount of clinical conditions found in SNOMED-CT. However, the Semantic Web enables access to the ontologies, such as SNOMED, via OWL-based formalization to promote interoperability.

Ontologies and language are related such that both draw on computational linguistics and knowledge engineering. Besides its application in semantic technologies, ontologies have similar functionality to lexicons. However, a lexicon is not an ontology. A lexicon consists of a list of words and sometimes accompanied by the words usage in a language, organized as an inventory. On the contrary, ontology formalizes concepts and their logical relations in a machine understandable way. Semantic lexicons are more focused towards capturing near-synonymy relations, whereas a formal ontology will group concepts as classes or subclasses under formal relationships such as 'is-a' or 'part-of'. The relation between ontologies and lexicons is bidirectional, since ontologies can enhance lexicons and vice-versa. An interesting direction of research is towards integrating the lexical resources with semantic technologies, more specifically to build ontolexical models, for improving the performance of NLP. However, the interrelationships between language and concepts are complex which introduce some challenges. Natural language exposes ambiguity and variation in expressing semantic behavior in the form of polysemy, metaphor, metonymy and vagueness. Contextual information heavily affects semantic meanings in texts, and hence the question arise at which stage should the ontolexical model be applied for NLP tasks - lexicon or ontology or at processing level. Having a model that contains both lexical and semantics seems to be promising, however the challenge lies in better understanding the relations that exists between concepts, lexical items and linguistic contexts.

Ontology learning is the process of acquiring knowledge from texts for ontology development. Similar to computational linguistics, ontology learning aims at (semi-)automatically retrieve lexical information, more specifically conceptual knowledge, from clinical notes. A number of methods are used for building ontologies, namely machine learning, knowledge acquisition, natural language processing, information retrieval, artificial intelligence, reasoning and databases.

One method of constructing ontologies is through learning methods from unstructured text. Natural language processing and statistical approaches are the most commonly used for acquiring knowledge from texts. Ryu and Choi, used term specificity and similarity to automatically build an ontology based on taxonomy (Ryu and Choi, 2006). The authors introduced the distributional hypothesis to extract taxonomy of terms. They argue that obtaining high precision from automatic taxonomic relation learning from unstructured texts is hard and human intervention is inevitable. Another system, where both NLP and statistical characteristics were retrieved from texts for ontology learning, was developed by Sanchez and Moreno (Sanchez and Moreno, 2004). Frequency counts of noun and noun phrases were used to discover concepts and taxonomic relations from the web. Text2Onto is another framework for ontology learning which uses an integrated approach from information retrieval, lexical databases, machine learning and

computational linguistics (Cimiano, 2005). Another approach is the Formal Concept Analysis (FCA), which is based on lattice theory (Jiang et al., 2003). FCA is effective at automatic construction of formal ontologies for representing taxonomy relationships, such as is-a and equivalence. This mathematical analysis technique helps at making a partial-ordering relation between concepts and attributes, which may need further human refinement. Jiang et al. constructed a context-based ontology for a clinical domain and their results have been useful in supporting clinicians for ontology building tasks. However, the work performed in ontology learning so far, has been limited to binary relations and specific domains. Moreover, limited research has been performed in integrating linguistic and context-based information towards ontology learning. Therefore, challenges remain for novel methodologies that combine conceptual, linguistic and contextual information from clinical text. Likewise, in order to facilitate interoperability and reuse of IE systems, approaches that can adapt to new domains will be needed.

Besides using NLP techniques to extend existing human-made ontologies, recently there has been some motivation towards combining other resources for knowledge discovery. Liu et al. performed a study to investigate semantic space of WordNet and UMLS in the clinical domain corpora. Besides having some overlapping concepts from both, each resource also consisted of independent concepts. The authors deduce that each resource can contribute to identifying new concepts and combining general English with domain specific resources seem to be promising for natural language processing tasks (Liu et al., 2012b). Wikipedia is another opening towards adding more information or improving IE systems. In the clinical domain, MESH and SNOMED-CT are large domain resources and by incorporating these resources with clinical data, via network structure, provide opportunities for finding new concepts. Bate et al., implemented a new measure to find similar concepts from clinical corpora using ontology-based methods. The study showed that their measure outperformed most of the path-based and context vectors approaches (Batet et al., 2011). Moreover, the ontology-based methods open doors as an innovative pathway for integrating domain knowledge into machine learning techniques.

As mentioned above, ontologies are limited with respect to coarse grain lexical information. Term meaning can be distinguished by its different senses, inferences and similarity with other words. The ontology mainly keeps the term concept and its formal relationships, but not the variations in meaning. Therefore an ontology can be enriched by mapping terms to semantic classes, as well as representing the semantic context of the term under consideration to represent its lexical information. Ontology-based semantic lexicon is now possible, with the creation of lexicon models, like LingInfo, LexOnto and LexInfo (Buitelaar et al., 2009). LingInfo is an ontology model for representing linguistic information, such as inflection and morphosyntactic decomposition, while LexOnto enables term representations of predicate-argument structure. However, LexInfo tries to merge both representation of the previous two models enhanced with the Lexical Markup Framework (LMF) in order to have a more complete lexical representation. The LMF is an ISO approved standard for natural language processing and machine-readable dictionary. Reiter and Buitelaar used WordNet synsets metrics to populate the Foundational Model of Anatomy (FMA) ontology with lexical entries from a corpus of Wikipedia pages on human anatomy (Reiter and Buitelaar, 2008). Their approach resulted in retrieving significant lexical information on human anatomy, however they intend to use LingInfo model for building an FMA ontology-based lexicon representation. Moreover, a resource, such as FrameNet, is based on semantics that are now available in Web ontology Language-Description Logic (OWL-DL), which favors Description Logic reasoning. This promotes the use of standardized methods to integrate ontologies and lexicons. Ultimately, it promises interoperability through the Semantic Web theory to enable automatic extraction of semantics from text.

5 METHODOLOGY

The main objective of my thesis proposal is to leverage semantic lexicons for facilitating content interoperability for information extraction from clinical notes. The features that stimulate knowledge and ontologies are greatly influenced by contexts of their use. Knowledge is captured as entities and attributes in ontologies, however its usage dynamically varies and the ontologies are not sufficient for NLP tasks. In the clinical domain, UMLS and SNOMED-CT are the most widely used terminologies for IE or NLP tasks, such as phenotyping and cohort identification for secondary use. Besides the hierarchical complexity of the existing clinical terminologies, information from clinical notes is not completely covered by these terminologies. Moreover, the limited access to clinical corpora for research has hindered the progress towards clinical knowledge discovery. Yet independent lexicons for specific clinical IE tasks are being con-

structed which do not favor re-use. Novel methods and ontolexical models should be devised via adaptive IE for easing integration across sources. These would in turn promote secondary use of clinical data for improving healthcare and quality. The proposed research plan is organized around the following research objectives:

5.1 Identification of Lexico-semantic Relations and Patterns from Clinical Corpora

Addressing this objective begins with an analytical study of clinical notes content to understand the information structures, relations and patterns across clinical entities. The clinical domain consists of complex phrases that need to be understood through its components in order to be able to represent their meanings. Different approaches will be considered here in order to predict prevalent patterns from a clinical corpus, which is representative from different providers, sources and hospitals from various VISNs in the national VA healthcare system. The primary contributions to the first objective are procedures for:

A. Discovering and understanding conceptual relations

The emerging approach, namely Formal Concept Analysis (FCA), has proved to be useful for understanding conceptual relations in data, since it provides a systematic layout to represent formal contexts through lattices. Besides ensuring domain knowledge completeness, it has also been established in real-life practice for knowledge discovery with real outcome (Poelmans and Kuznetsov, 2013). FCA techniques will be used for knowledge discovery from clinical corpus. The following is an outline of the procedure that will be considered for relation discovery from clinical notes.

- Identifying clinical concepts as objects. Here, concepts will be distinguished by exploring existing terminologies, such as UMLS and SNOMED-CT, and documentation of specific knowledge.
- Identifying the most appropriate set of attributes to describe the terms. Attribute selection algorithms, will be investigated by exploring word types and lexico-syntactic contexts.
- Extracting the implied relations between the objects and attributes. This will involve analysis of the conceptual structures by using lattice theory and clustering techniques.

B. Identifying semantic relations among entities

This task will focus on discovering significant relations automatically between entities from clinical documents. We will investigate feature extraction, selection methods and clustering algorithms. The approach is to use clustering techniques to group frequently co-occurring entities that appear in similar contexts. In contrast to part A. above, which captures hierarchical binary relations between two entities, in this process the focus will be towards capturing more generic relations based on empirical distributional characteristics from corpora. The following describe the procedures that will be required:

- Finding a set of candidate relationships, by finding entities or predicate-argument structures that frequently co-occur in the same sentences. Clinical terminologies will be exploited for initial concept and relationship mappings. Then, the possible approaches may be by finding kernels or concepts with modifiers of the verb, modifiers of its arguments and connection with other kernels will be investigated for finding candidate relationships.
- Sentences having similar meanings that are described in different contexts will be analyzed and clustered together. Here distributional modeling techniques will be used to compare contextual word distribution across corpora.
- The significant clusters that exhibit repetitive patterns of word choice will be captured as sublanguage word classes and sublanguage sentence structures.

5.2 Building an Adaptive Information Extraction Application

In this proposal, the relation identification and actual relation extraction tasks are separated. Therefore, this task will include the relation extraction task by using the frequently used high-precision features from the previous aim. In order to balance the precision, recall will be gained by extracting as many relation instances. This methodology was introduced by Shinyama and Sekine [75]. The motivation is to reduce time and human effort for discovering all possible repeated relations automatically from a corpus to create feature tables. Therefore, the features or patterns identified from the previous aims will be used to extract associated lexical terms using a learning algorithm.

The extracted relations will need to be added to

the semantic lexicon and tied to the correct concept. In order to find the closest concept that the new term should belong to, its similarity will need to be determined and existing ontologies will be used. The graph based modeling algorithms seem to be an intuitive methodology for the lexical acquisition tasks. An ensemble of graph modeling and distributional analysis techniques has been used by (Widdows and B, 2002) for extracting terms having similar meanings and also identifying terms that may have multiple meanings. Similarly, Wu and Liu (Wu and Liu, 2011) linked clinical name entities found in the text to SNOMED. Then the hierarchical propagation was used to compute the frequency counts of concepts for each domain. As a result a weighted ontology describing the distributional semantic information was developed. The authors supports the fact that this resource could help to improve NLP performance if applied to statistical processing and machine learning for information extraction tasks.

The UMLS has 12 different types of hierarchical and non-hierarchical relations, and standard ontology methods will require extensive adaptation. However, the SNOMED-CT is widely used and is formalized into description logic that creates opportunities for automated reasoning. Moreover, according to Kate et al., SNOMED-CT is an extensive resource and the expressions are described in terms of relations with other SNOMED-CT concepts (Kate, 2013). This creates opportunities to find new concepts by identifying relations between clinical phrases and SNOMED-CT concepts. Further, SNOMED allows multiple inheritances, which in turn provide multiple possibilities to explore semantic similarity of any two concepts. Therefore, for this task, graph modeling learning algorithms using clinical corpora, the SNOMED-CT terminology and additional lexical resources, such as Wikipedia, WordNet and VerbNet, with machine learning methods will be investigated to identify new relations or expressions.

5.3 Construction of an Ontology-based Semantic Lexicon using Semantic Web Technologies and Lexical Ontologies

This aim will consist mainly of modeling of an ontology-based semantic lexicon, focusing on formalization, consistency and completeness of the ontology. In order to integrate lexical resources with ontologies, two approaches can be taken. In both approaches the ontology and lexical resources are kept independent. The first direction is to start from an ontological framework and then enhance the ontology with knowledge structure of lexical resources. Chou and Huang built an ontology based on semantic-based orthographic system of Chinese language. They adopted Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2001) as representational framework for mapping the Chinese characters concepts. The Chinese characters may contain multiple meanings depending on the attachment of the character with other characters. Therefore, they proposed to use linguistic context as the next step to describe character variants (Chou and Huang, 2010). OWL-DL was also used to make the information computable and sharable. The other direction is to start from the lexical resources to populate lexical terms into ontological knowledge. A typical example is to construct a domain ontology based on taxonomical, such as hyponym-hypernym and non-taxonomical, such as synonyms, meronyms and functional relations.

Therefore, the approach will be to classify lexical units according to semantic classes. Different ontological framework, such as SUMO for structural ontology representation and LexInfo for lexical structure representation, will be investigated. As described in the literature review above, LexInfo provide supplemental linguistic structure, such as PoS, predicate-argument, among others. The ultimate goal will be to investigate how both conceptual and lexical information could be represented in a formal way. Both frameworks support the OWL-DL and the advantage of using such an encoding representation will render the proposed ontolexical model as a sharable resource that could be used for SemanticWeb applications in the future. Moreover, OWL-DL provide opportunities for automatic reasoning and easy integration with linguistic vocabulary, such as VerbNet and FrameNet. Further, adopting the SUMO representation, provide integration opportunities with linguistic ontologies such as WordNet.

6 EXPECTED OUTCOME

This research study will contribute to term discovery and methods to integrate usage information from clinical narratives. Information structures derived from clinical text will be considered to resolve term-concepts ambiguities. Methods to leverage semantic lexicons for clinical information extraction will intensify reconciliation of structured and unstructured data in Electronic Health Records thus reducing the amount of false positive records in patient cohort identification processes. The following outcomes are expected:

- Identification of features and sublanguage structures that are associated to clinical concepts, such as homelessness and psycho-social factors.
- Discovery of phrases and terms associated to the clinical domain.
- Minimizing supervised learning in lexicon building.
- Provide a standardized ontology based semantic lexicon for interoperability and improved NLP performance.
- Allow re-use of the ontology based semantic lexicon in more than one natural language processing system.
- Enabling maintainability, where new items could be added to the semantic lexicon in a consistent manner.

In addition to promote semantic interoperability for NLP applications, this research study has also revealed practical implications for the VA. A suggestion for future study would be to determine accurate estimates of patients from secondary use of EHR data. For instance, a practical implication would be to study the extent of homeless Veterans among different VISNS and refined estimates of homelessness could be established. Above all, the proposed methodology could be an opportunity to standardize the terminology related to homelessness in different VA medical facilities around the country.

REFERENCES

Abacha, A. and Zweigenbaum, P. (2011). Medical entity recognition: A comparison of semantic and statistical methods. *Proceedings of BioNLP 2011 Workshop*, pages 56–64.

Batet, M., Sánchez, D., and Valls, A. (2011). An ontology-based measure to compute semantic similarity in biomedicine. *J Biomed Inform*, 44(1):118–25.

Birman-Deych, E., Waterman, A. D., Yan, Y., Nilasena, D. S., Radford, M. J., and Gage, B. F. (2005). Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Medical care*, 43(5):480–5.

Buitelaar, P., Cimiano, P., Haase, P., and Sintek, M. (2009). Towards linguistically grounded ontologies. *The Semantic Web: Research and Applications Lecture Notes in Computer Science*, 5554:111–125.

Chou, Y.-M. and Huang, C.-R. (2010). *Hantology: conceptual system discovery based on orthographic convention*. CAMBRIDGE University Press.

Cimiano, P. (2005). Text2onto–a framework for ontology learning and data-driven change discovery. *NLDB'05 Proc 10th Int Conf Nat Lang Process Inf Syst*.

Coden, A. R., Pakhomov, S. V., Ando, R. K., Duffy, P. H., and Chute, C. G. (2005). Domain-specific language models and lexicons for tagging. *J Biomed Inform*, 38(6):422–30.

Cohen, A. M., Hersh, W. R., Dubay, C., and Spackman, K. (2005). Using co-occurrence network structure to extract synonymous gene and protein names from medline abstracts. *BMC Bioinformatics*, 6:103.

Friedlin, J. and Overhage, M. (2011). An evaluation of the umls in representing corpus derived clinical concepts. *AMIA Annu Symp Proc*, 2011:435–44.

Friedman, C., Kra, P., and Rzhetsky, A. (2002). Two biomedical sublanguages: a description based on the theories of zellig harris. *Journal of Biomedical Informatics*, 35(4):222–235.

Friedman, C., Rindflesch, T. C., and Corn, M. (2013). Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *Journal of biomedical informatics*, 46(5):765–73.

Grishman, R. (2001). Adaptive information extraction and sublanguage analysis. *Proc. of IJCAI 2001*, pages 1–4.

Guthrie, L., Pustejovsky, J., Wilks, Y., and Slator, B. M. (1996). The role of lexicons in natural language processing. *Commun. ACM*, 39(1):63–72.

Howes, C., Purver, M., and McCabe, R. (2013). Investigating topic modelling for therapy dialogue analysis. *WCS 2013 Workshop on Computational Semantics in Clinical Text*.

Huang, J., Dou, D., Dang, J., Pardue, J. H., Qin, X., Huan, J., Gerthoffer, W. T., and Tan, M. (2012). Knowledge acquisition, semantic text mining, and security risks in health and biomedical informatics. *World J Biol Chem*, 3(2):27–33.

Jiang, G., Ogasawara, K., Endoh, A., and Sakurai, T. (2003). Context-based ontology building support in clinical domains using formal concept analysis. *Int J Med Inform*, 71(1):71–81.

Johnson, S. B. (1999). A semantic lexicon for medical language processing. *J Am Med Inform Assoc*, 6(3):205–18.

Jonnalagadda, S., Cohen, T., Wu, S., Liu, H., and Gonzalez, G. (2013). Using empirically constructed lexical resources for named entity recognition. *Biomed Inform Insights*, 6(Suppl 1):17–27.

Kate, R. J. (2013). Towards converting clinical phrases into snomed ct expressions. *Biomed Inform Insights*, 6(Suppl 1):29–37.

Liu, H., Wu, S. T., Li, D., Jonnalagadda, S., Sohn, S., Wagholikar, K., Haug, P. J., Huff, S. M., and Chute, C. G. (2012a). Towards a semantic lexicon for clinical natural language processing. *AMIA Annu Symp Proc*, 2012:568–76.

Liu, Y., McInnes, B. T., Pedersen, T., Melton-Meaux, G., and Pakhomov, S. (2012b). Semantic relatedness study using second order co-occurrence vectors computed from biomedical corpora, umls and wordnet. *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI '12*.

MacLean, D. L. and Heer, J. (2013). Identifying medical terms in patient-authored text: a crowdsourcing-based approach. *J Am Med Inform Assoc*, 20(6):1120–7.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., and Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, pages 128–44.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. *Proceedings of the international conference on Formal Ontology in Information Systems - FOIS '01*, pages 2–9.

Patterson, O. and Hurdle, J. F. (2011). Document clustering of clinical narratives: a systematic study of clinical sublanguages. *AMIA Annu Symp Proc*, 2011:1099–107.

Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., and Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *J Biomed Inform*, 40(3):288–99.

Poelmans, J. and Kuznetsov, S. (2013). Formal concept analysis in knowledge processing: A survey on models and techniques. *Expert Systems with Applications*, pages 1–44.

Reiter, N. and Buitelaar, P. (2008). Lexical enrichment of a human anatomy ontology using wordnet. *Proc. Global WordNet Conference (GWC)*.

Ryu, P. and Choi, K. (2006). Taxonomy learning using term specificity and similarity. *Proceedings of the 2nd Workshop on Ontology Learning and Population. Association for Computaional Linguisitcs*, pages 41–48.

Sager, N., Lyman, M., Nhàn, N. T., and Tick, L. J. (1995). Medical language processing: applications to patient data representation and automatic encoding. *Methods Inf Med*, 34(1-2):140–6.

Sanchez, D. and Moreno, A. (2004). Creating ontologies from web documents. *Recent Adv Artif Intell Res Dev 2004; IOS Press*.

Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P. J., Elhadad, N., Johnson, S. B., and Lai, A. M. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 21(2):221–30.

Sohn, S., Clark, C., Halgrim, S. R., Murphy, S. P., Jonnalagadda, S. R., Wagholikar, K. B., Wu, S. T., Chute, C. G., and Liu, H. (2013). Analysis of cross-institutional medication description patterns in clinical narratives. *Biomed Inform Insights*, 6(Suppl 1):7–16.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–6.

Verspoor, K. (2005). Towards a semantic lexicon for biological language processing. *Comp Funct Genomics*, 6(1-2):61–6.

Widdows, D. and B, D. (2002). A graph model for unsupervised lexical acquisition. *COLING '02 Proc 19th Int Conf Comput Linguist*, 1:1–7.

Wu, S. (2013). Computational semantics in clinical text. *Biomed Inform Insights*, 6(Suppl 1):3–5.

Wu, S. and Liu, H. (2011). Semantic characteristics of nlp-extracted concepts in clinical notes vs. biomedical literature. *AMIA Annu Symp Proc*, 2011:1550–8.

Wu, S. T., Liu, H., Li, D., Tao, C., Musen, M. A., Chute, C. G., and Shah, N. H. (2012). Unified medical language system term occurrences in clinical notes: a large-scale corpus analysis. *J Am Med Inform Assoc*, 19(e1):e149–56.

Zhang, Z., Gentile, A., Xia, L., Iria, J., and Chapman, S. (2010). A random graph walk based approach to computing semantic relatedness using knowledge from wikipedia. *LREC*, pages 1394–1401.

Zweigenbaum, P., Lavergne, T., Grabar, N., Hamon, T., Rosset, S., and Grouin, C. (2013). Combining an expert-based medical entity recognizer to a machine-learning system: methods and a case study. *Biomed Inform Insights*, 6(Suppl 1):51–62.