Exploiting Semantic Structure for Mapping User-specified Form Terms to SNOMED CT Concepts

Ritu Khare
ritu@ischool.drexel.edu

Yuan An
yan@ischool.drexel.edu

Jiexun Li
jiexun.li@ischool.drexel.edu

Il-Yeol Song
isong@ischool.drexel.edu

Xiaohua Hu
thu@ischool.drexel.edu

The iSchool at Drexel
Drexel University
Philadelphia, PA-19104, USA

ABSTRACT

The elements of clinical databases are usually named after the clinical terms used in various design artifacts. These terms are instinctively supplied by the users, and hence, different users often use different terms to describe the same clinical concept. This term diversity makes future database integration and analysis a huge challenge. In this paper, we study the problem of standardization of the terms used in a specific kind of user-designed artifact, *the encounter forms or templates*, using a popular clinical terminology, *the SNOMED CT*.

In particular, we focus on the problem of mapping the terms on an encounter form to SNOMED CT concepts. Existing term mapping techniques are solely based on syntactic string similarity. Such techniques are unable to disambiguate among the terms that resemble one another linguistically, and yet differ semantically. To improve existing techniques, we consider the context of a term in the mapping process and propose a hybrid approach relying on linguistics as well as structural information. For a given form term, this approach (i) exploits the semantic structure of the form to derive the term's context, and (ii) maps the term to a linguistically- matching SNOMED CT concept that is compatible with the derived context. We test the approach on over 900 clinician-specified terms used in 26 forms. This method achieves 23% improvement in precision and 38% improvement in recall, over a pure linguistic-based approach. Our first contribution is that we introduce and address a new problem of mapping form terms to standard concepts. The second contribution is that the experimental evaluation confirms that structural information has a major role in improving mapping performance, and in addressing the key challenges associated with semantic mapping.

IHI'12, January 28–30, 2012, Miami, Florida, USA.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval—*Information Search and Retrieval*; J.3 [**Computer Applications**]: Life and Medical Systems

General Terms

Algorithms, Experimentation, Performance

Keywords

Forms, Mapping, Semantics, Structure, SNOMED CT, Terms

1. MOTIVATION

Semantic heterogeneity across clinical data sources makes database integration and interoperability a huge challenge[9, 12, 33, 11]. Heterogeneity is mainly caused by the diversity of the terms selected by users to design or populate different healthcare databases. To facilitate interoperability across disparate databases, it is important to incorporate controlled clinical terminologies into design artifacts including user interfaces and back-end databases [31, 14].

Clinical encounter forms are an important tool in electronic health record (EHR) systems for collecting data into databases. The terms on an encounter form are often specified by the user, and are directly associated with the elements in the underlying database schema and instances. It would greatly reduce the database heterogeneity if the terms on the clinical forms are mapped to, or annotated by, a standard terminology. Although a knowledge engineer can carefully design encounter forms and databases conforming to a standard terminology, this process is very costly and tedious. Also, there are other cases where either legacy systems need to be mapped to a standard terminology, or the non-technical users, e.g., clinicians, want to specify their own encounter forms. For these cases, it is desirable for an automatic tool to assist users in mapping form terms to standard terminologies.

In this work, we study the **problem of mapping terms of clinical encounter forms to SNOMED CT concepts**, and develop a context-based method that leverages the semantic structure of forms to improve the mapping results. The *Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT)* is a widely used medical terminology. It is comprised of over 360,000 logically-defined clinical

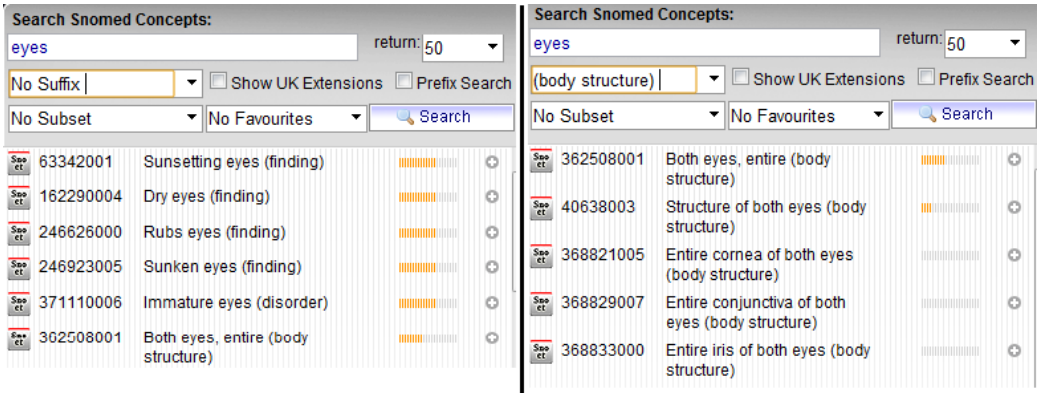

Figure 1: Mapping the term "eyes" to SNOMED CT. (a) general mapping (b) category-specific mapping

concepts belonging to various *semantic categories* [30, 13]. Each concept is represented using a numeric *concept id* and multiple kinds of *descriptions*. One kind of description is the *fully specified name* that ends with the semantic category label, e.g., the description "Ocular hypermia (disorder)" implies that the concept belongs to the semantic category, disorder. In addition, the concepts are related to each other by *defining relationships*.

Compared to the traditional schema and ontology mapping problem [27, 21, 5, 15, 32], the problem of mapping forms to SNOMED CT raises several new challenges. First, a form is graphical user interface that lacks a well-defined in-built semantic structure among the form elements. Automatic form understanding is challenging [17]. Second, the SNOMED CT is a large medical knowledge base that encodes concepts and relationships from many aspects of clinical information. Terminology navigation and efficient retrieval of relevant terms is difficult. Third, both forms and SNOMED CT usually do not have instances. Hence, the instance-based techniques for schema mappings are hard to apply. Finally, forms and the SNOMED CT are two entirely different structures. It is almost infeasible to convert them into a uniform formalism.

There are SNOMED CT terminology services that allow users to search concepts through the indices of the concept descriptions in the SNOMED CT. The key problems with the current systems include too many irrelevant results and inability to distinguish semantic categories. In this paper, we introduce and address the problem of mapping a given form term to a unique SNOMED CT concept. We focus on extracting the context of a term, and on using the context to improve the results of retrieving relevant SNOMED CT concepts.

1.1 Issues with Existing SNOMED CT Services

Let us first consider solving the mapping problem using existing services. There are several browsers that provide public access to the SNOMED CT[28]. Underneath these browsers, the user-supplied keyword is compared with the descriptions of SNOMED CT concepts using certain linguistic techniques, and a ranked list of all matching concepts is created and returned for browsing purposes. These services can be understood to provide two kinds of mapping: (i) general mapping, wherein the user term is matched against all the SNOMED CT concepts, (ii) category-specific mapping, wherein the term is matched only against the concepts belonging to a specific semantic category.

As an example, we consider the *Snoflake* browser provided by the Dataline Software Ltd[1]. To perform the mapping, *Snoflake* looks for the SNOMED CT descriptions that contain the search term, and returns the associated concepts for further browsing. Each concept is assigned a *match-weight*, which is calculated as the overlap ratio between the words contained in the search term and the words contained in the concept's fully-specified name. The concepts are sorted in a non-increasing order of their weights, and in an increasing order of their concept identifiers for equally weighing concepts. Figures 1a and 1b show the screen shots of the results of mapping the term "eyes" using general and category-specific methods, respectively. For each retrieved concept, the browser returns the concept id, the fully-specified name, and a visual bar representing the match-weight. Despite their public availability, these browsing services are inadequate to address the mapping problem due to the following reasons.

- Different clinicians specify different form terms to describe the same clinical concept, e.g., the use of abbreviations ("MRN," "Med. Rec.#") or synonymous and hyponymous terms ("vital signs," "constitutional," "physical status"). The services are not designed to handle the wide variation in the terms. We refer to this user-induced challenge as the *diversity challenge*.

- Another issue is due to the inherent richness of forms and SNOMED CT. The same form term, when used in different contexts, may map to different concepts. For instance, in Figure 2, the element labeled with the term "Respiratory" in Form 1 maps to a concept belonging to the body structure semantic category; another element labeled with the same term in Form 2 maps to a concept belonging to the observable entity category. This disambiguation task entails expert judgment. Moreover, a single term may linguistically match with multiple concepts, and locating the desired concept within this large result set also requires human intervention. We refer to this as the *context challenge*.

Both these challenges are different flavors of the long-standing ambiguity problem in the field of natural language processing[29]. While several linguistic-based works[11, 16,

26, 27] exist for addressing the diversity challenge, the context challenge for mapping form terms is not much explored. In this work, we focus on this challenge and propose a form structure-based approach to automatically retrieve an accurate concept corresponding to a given term.

1.2 Mapping Forms to SNOMED CT

The proposed approach is based on the following premises. First, the key to the mapping problem is to identify the SNOMED CT semantic category appropriate for a given term. Once this identification is done, the first, i.e., the most string-similar, result retrieved by the category-specific mapping is usually the desired concept. For example, consider the element labeled with the term "Eyes" from Form 1 in Figure 2. If there is a mechanism to determine its semantic category, which in this case is body structure, then the desired concept could be recovered through a category-specific mapping, as shown in Figure 1b. The second premise is that the identification of a term's semantic category requires the knowledge of the *context* in which the term has been specified. We hypothesize that the term context can be derived from the semantic structure of the form, and that the implicit relationship between the term context and the desired semantic category can be formally captured into a statistical model. To materialize this, we employ the `form tree`, a representation construct to capture the semantic structure of a form; and we devise a machine-learning based model, the `sClassifier`, that classifies a given term into a semantic category based on the structure of the `form tree`.

In sum, the proposed approach functions in the following manner. (1) Determine the SNOMED CT semantic category of a given form term using the structure-based model. (2) Perform a category-specific mapping and map the term to the first returned concept. This work makes the following contributions.

1. Despite the prevalence of the use of forms for populating and designing healthcare databases, standardization of form terms has not been studied in the past. We introduce this new problem of mapping user-specified form terms to SNOMED CT concepts. The solution exploits the semantic structure of forms to complement and improve the existing linguistic-based mapping approaches.

2. We evaluate the solution on 26 healthcare forms containing over 900 mappable terms. To quantify the performances, we use the popular information retrieval measures: precision and recall[22]. The proposed hybrid approach leverages both structure and linguistics, and achieves a 23% improvement in mapping precision and a 38% improvement in mapping recall, over an existing linguistic-based approach. The approach achieves a precision of up to 0.89, and a recall of up to 0.76.

3. We measure the individual impact of linguistics and structure on the mapping performance. We find that while the linguistic component helps address the diversity challenge and improves the mapping recall, the structural component helps address the context challenge and improves the precision as well as the recall. It is hence desirable to design synergistic hybrid approaches that have the potential to overcome the mapping challenges and achieve a high performance.

Section 2 of this paper introduces the representation schemes for forms and SNOMED CT concepts. Section 3 describes the framework to map the form terms to SNOMED CT concepts. Subsection 3.1 presents the flexible electronic health record (fEHR), a specific application of the proposed approach. Subsection 3.2 presents the solution to the mapping problem. Section 4 describes the mapping experiments conducted using real-world clinician-designed forms. Section 5 summarizes the related work. Finally, Section 6 concludes the paper.

2. SEMANTIC REPRESENTATIONS

The mapping problem is about finding semantically, and not just linguistically, matching SNOMED CT concepts for form terms. For this, we need the schemes to accurately capture the semantics of forms as well as SNOMED CT. This section describes the representation schemes adopted for both.

2.1 Forms

Encounter forms are designed primarily for data collection. Figure 2 shows two example forms. A form is a logically organized collection of form elements where each element is either a `text-label`, e.g., *PATIENT*, *Name*, *M*, or a `form-input` such as textbox, textarea, radiobutton, checkbox, etc. What is noteworthy is that a form is not just a thoughtful arrangement of elements to facilitate data-entry; it also reflects the user's view of the semantic associations among the elements, e.g., the parent-child associations such as *PATIENT-Name*, and sibling associations such as *PATIENT-HISTORY*.

A form could be represented using multiple schemes. The simplest is the source code itself, which is an ordered sequence of the form elements. Following this convention, the Form 1 in Figure 2 is represented as <*Patient*, *Name*, textbox, *DOB*, textbox,...>. However, such a flat representation fails to capture the designer's precise intentions, i.e., the semantic associations among the elements. Another way is to represent the form as a syntactic DOM tree[8]. Counterintuitively, even this representation fails to capture the semantic parent-child associations among the elements. The DOM tree is necessarily a syntactic tree of the formatting elements in a specific language, e.g., HTML tags <FONT>, <PARA>, etc.; such representations capture no information on the semantic grouping of the form elements.

In the context of the semantic mapping problem, we employ a new representation scheme known as the `form tree` that accurately captures the designer's intentions, and hence the semantic associations among the form elements. Definition 2.1 formally describes the form tree.

DEFINITION 2.1 (FORM TREE). *A Form Tree is defined as a labeled, directed, and ordered tree, $\mathcal{FT} = (N, E, root)$, where*

- *N is a finite set of nodes. Each $n \in N$ has a label l and a type t. Each node is of one of the following types:* label *node, representing a text label on the form;* format *node {*textarea, textbox, radiobutton, checkbox, dropdown*}; or* value *node. The function $\lambda(n)$ returns the node label, and the function $\tau(n)$ returns the node type.*

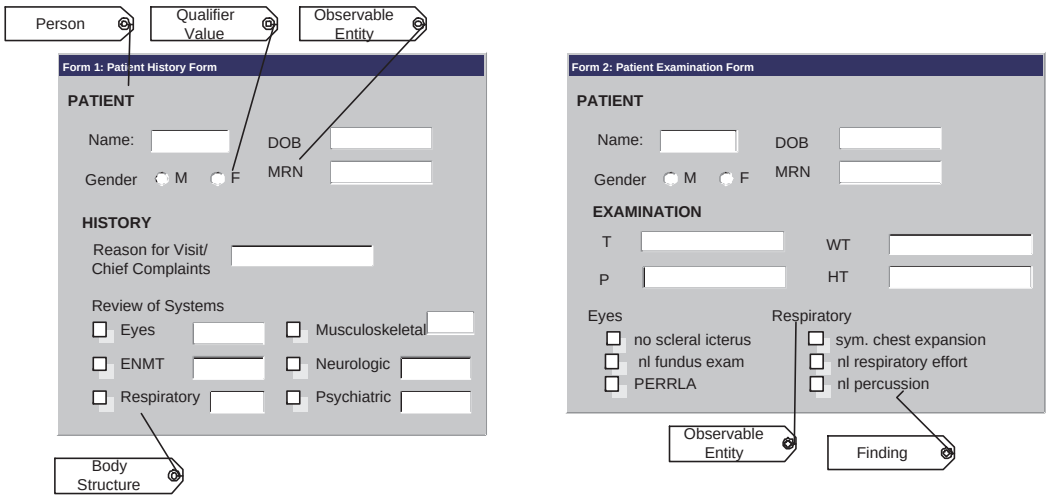

Figure 2: User designed Forms. Tags represent the SNOMED CT semantic categories

- *E is a finite set of edges, such that for an edge* $(n_i, n_j) \in E$, $n_i, n_j \in N$, n_i *is called* parent *and* n_j *is called* child.
- $root \in N$ *is the root node of the tree.*

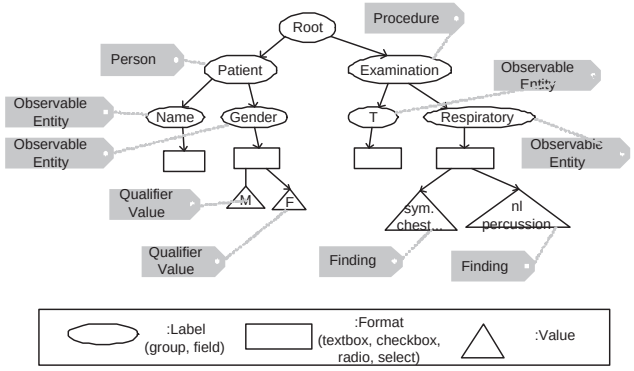

Figure 3: A Form Tree and the associated SNOMED CT semantic categories

Figure 3 shows a partial `form tree` corresponding to the Form 2 from Figure 2. A `form tree` could be extracted from a given form by an automatic method, such as the sophisticated machine learning technique presented in [4], or could be simultaneously derived while the user is creating a form, as in the tree extraction module of the fEHR system[18], which captures the intentions of the user on-the-fly. In this paper, we assume that a `form tree` is generated by one of these techniques when a form is given or created.

2.2 SNOMED CT

The SNOMED CT services are owned, maintained, and distributed by the International Health Terminology Standards Development Organization[2]. The SNOMED CT is the most comprehensive clinical vocabulary that precisely represents clinical information across the scope of healthcare[3]. It consists of *concepts*, *terms*, and *relationships*. Each *concept* is identified by a unique identifier, *concept id*, and is represented by a unique human readable term known as the *fully specified name*.

Table 1: Descriptions of a SNOMED CT Concept

Description	Value
Fully Specified Name	Respiratory rate (observable entity)
Preferred Term	Respiratory rate
Synonym	Rate of Respiration
Synonym	Respiration Frequency

Table 2: SNOMED CT Relationships

Relationship Type	Related Concept
Is a	Bone Injury(disorder)
Associated morphology	Fracture(morph. abnormality)
Finding Site	Bone Structure(body structure)

A concept is associated with multiple *descriptions*, which are the terms used to describe the concept. Each concept has 3 kinds of descriptions: (i) fully specified name: an unambiguous way to name a concept, (ii) preferred term: the most common term used by the clinicians to describe the concept, and (iii) synonym: additional terms used to describe the concept. As an example, the descriptions of a concept (concept id: C0231832), are enlisted in Table 1. The fully specified name ends with a parenthesized text that represents the *semantic category* to which the concept belongs, e.g., observable entity in this case. There are 19 semantic categories in SNOMED CT. The concepts are associated with each other using defining *relationships*. For example, the relationships of the concept *Fracture of bone (disorder)* with other concepts are enlisted in Table 2.

In this work, we are interested in concepts, fully specified names, and semantic categories. The semantic category of a given concept represents the top-level granularity concept associated with the concept through the IS-A relationship.

3. MAPPING TERMS TO CONCEPTS

In this section, we present our solution to map a form term to a SNOMED CT concept. The target application of the

mapping approach is any methodology that employs forms to design and/or populate databases, wherein the user supplied form terms are used for naming the database elements. Through mapping, we intend to standardize the user terms to generate standard annotated databases, and thereby support future data integration and analysis.

We have shown that solutions solely based on the linguistic similarity between the term and the concept description do not achieve good accuracy. This is because a form term does not stand alone and is strongly associated with a certain context within the form. The same term when used in different contexts maps to different SNOMED CT concepts from different semantic categories. To address this, we propose a solution that (i) exploits the semantic structure of forms to determine the context, and the appropriate semantic category for a given term, and (ii) maps the term to a linguistically matching concept within the determined semantic category. Before we present our solution, we present the fEHR system, one of our initial motivations to devise an approach to map forms to SNOMED CT.

3.1 fEHR: A Target Application

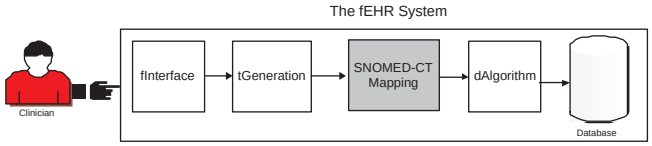

Figure 4: The flexible EHR System with the new (shaded) SNOMED CT Mapping component.

The fEHR system is a flexible system that allows the clinicians to easily extend the data collection functionality based on their customized needs. This is made possible by incorporating a form-driven approach. The system leverages the facts that the knowledge of forms is already ingrained in clinicians and the semantics of forms could guide the generation of a database. Figure 4, minus the shaded component, shows the fEHR architecture. The system has 3 components: (i) fInterface that allows clinicians to self-design need-based forms, (ii) tGeneration that derives the semantic form tree for the input form, (iii) dAlgorithm that explores the tree structure to build a high quality [7] database. For details, please refer to our previous work [18].

The form terms specified by the clinicians are eventually used to name the elements of the database schema, e.g., the term "Patient," used in the forms in Figure 2 is likely to be used to name a database table. At present, the terminology process in the fEHR system is uncontrolled in that the clinicians are free to supply any terms to the system. Due to differences in perceptions and domain expertise, it is highly likely that different clinicians would specify the same concept in different ways, thereby causing complications in future integration and analysis. To address this concern, we add a new middleware component to the fEHR system that maps the user defined form terms into SNOMED CT concepts, thereby generating standard-annotated database schemas. While fEHR is a specific application, the proposed mapping approach can be employed by any application that utilizes forms to design and populate clinical information systems.

3.2 The Mapping Solution

As discussed in Section 1, there are two main challenges associated with the mapping problem: the diversity challenge and the context challenge. An intuitive solution is to linguistically match the form term with the SNOMED CT concept descriptions and return the most matching concept. An example of such a linguistic technique is the general mapping provided by any SNOMED CT browser as shown in Figure 1a. Such methods, when combined with sophisticated term processing techniques, can certainly address the diversity challenge to a great extent. However, such methods treat every term as a context-independent entity. As such, they fail to disambiguate the context in which the term has been specified, and to determine the appropriate SNOMED CT semantic category for a given term. As a result, the context challenge remains unaddressed.

To address this special challenge, an advanced simulation of the category-specific mapping (See Figure 1b) is needed, that automatically determines the semantic category for a given term based on its context, and maps the term to a linguistically matching concept belonging to that category. The key in performing this simulation is the automatic identification of the semantic category based on the context of the form term. To accomplish this, we design a statistical model that exploits the structure of the semantic `form tree` to derive the term context and predict the SNOMED CT semantic category. In the next subsections, we first describe this structure-based model, and then illustrate the overall approach.

3.2.1 Structure-based Model

As mentioned before, we believe that the key to the mapping solution is to determine the semantic category appropriate for a given form term. What we intend to achieve is depicted in Figure 3 that shows a `form tree`, and the semantic categories associated with the terms contained in various nodes. A human expert can intuitively determine the category by perceiving the context of a term as implied in the form. To accomplish this automatically, we hypothesize that the context of a term contained in a given node can be extracted from the structure of the form tree. We encode the intuitive expert knowledge into a statistical model that earns the ability to determine the semantic category of any form term based on the tree structure. We refer to this model as the `sClassifier`. Following is a technical description of the model.

We use the Naïve Bayes Classifier [10] to design the `sClassifier`. The Naïve Bayes Classifier is one of the most effective classifiers based on the powerful Bayes theorem that determines the posterior probability, $P(\mathcal{H}|X)$, that the hypothesis $\mathcal{H}$ holds for an observed data sample X. Each sample is represented as $X = (x_1, x_2, ...x_n)$ depicting measurements from n attributes $A_1, A_2, ..., A_n$. For a given set of m classes, $C_1, C_2..., C_m$, the classifier calculates the posterior probability $P(C_i|X)$ for each class and assigns the data sample to the class with the maximum value. This classifier works with an assumption of class conditional independence, which states that the effect of an attribute on a given class is independent of the values of other attributes. To customize `sClassifier` for the problem of classifying a form term into a SNOMED CT semantic category, we use the following parameters.

***Class Labels*:** The classes comprise the pre-defined SNOMED

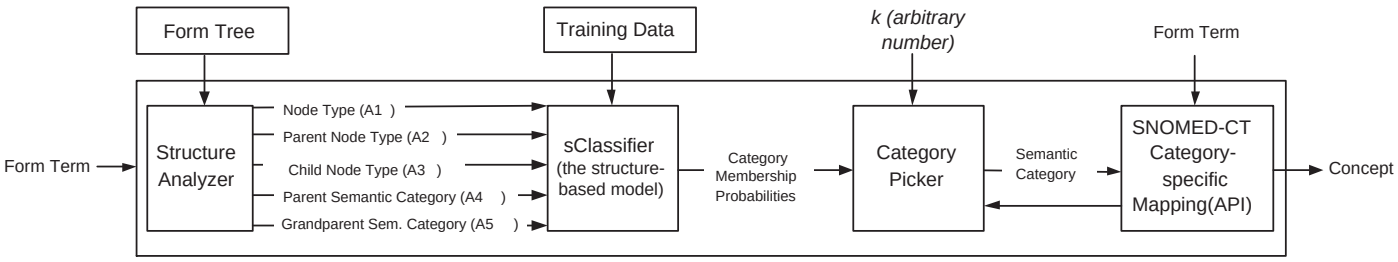

Figure 5: SNOMED CT Mapping

CT semantic categories. Out of all the available semantic categories, we choose the ones that are frequently associated with clinical form terms. In particular, the model employs the following class labels: attribute, body structure, disorder, finding, observable entity, occupation, person, physical object, procedure, product, qualifier value, racial group, record artifact, and situation.

Data Attributes: Given any term, $\lambda(n)$, contained in a node n, the goal of the classification process is to predict the most appropriate class label based on the term context. This implies that the classification attributes should reflect the context of the node n that holds the term. We hypothesize that the node context can be extracted from the local structure in the semantic form tree, and choose the following categorical attributes to accomplish classification:

1. Node type ($\tau(n)$): As per the Definition 2.1.
2. Parent node type ($\tau(n_j)$): The node type of the parent node n_j of the node n.
3. Child node Type ($\tau(n_i)$): The node type of the first child node n_i of the node n.
4. Parent semantic category: The semantic category of the parent node n_j, as determined by the `sClassifier`.
5. Grandparent semantic category: The semantic category of the grandparent node n_k, as determined by the `sClassifier`.

The domain of the values taken by the first three attributes includes label(group, field), field format (textbox/ checkbox/ radiobutton/ select), and value as defined in Definition 2.1. The domain of values taken by the last two attributes is same as that of the class labels. As an example, the values for the 5 attributes for the node labeled as "T" in Figure 3 are *field label*, *group label*, *textbox*, procedure, and *null* (since the tree root is not associated with any semantic category), respectively.

The main goal of this work is to study the impact, of exploiting the form structure, on the mapping performance. To create the structure-based model, we experiment with the Naive Bayes Classifier. In the future, we also intend to study the impact, of using other classifiers such as k Neural Networks and Classification Association Rules, on the model's performance.

3.2.2 Overall Approach

The overall approach to find a unique SNOMED CT concept suitable for a given form term is summarized in Figure 5. The approach is hybrid in nature, in that the first 3 modules are structure-based, and the last one is linguistic-based. The last module could be any application programming interface (API) that provides programmatic access to search and browse the SNOMED CT based on certain linguistic techniques.

The input to the mapping approach is the form term, contained by a particular node in the form tree. The first module, structure analyzer, exploits the structure of the form tree to extract the context of the term. The context, represented as the 5 attributes, A_1, A_2, A_3, A_4, A_5, is fed as the input to the structure-based model, the `sClassifier`. This trained model determines the class membership probabilities for the given term. In other words, the model determines the probability that the term belongs to any of the 15 semantic categories.

The next module, category picker, sorts the probabilities in the non-increasing order of values. It then picks the top ranked category, and performs the "concept presence test" using the API module. The test determines whether any SNOMED CT concept, with a linguistic match between the term and the descriptions, exists in the given category. If the test is positive, then the control is passed over to the API module that performs a category-specific mapping and returns the "most" linguistically matching concept as the output. However, the test result may also be negative mainly because: (i) the training data may be inconsistent; the terms having the same attribute values may belong to different classes, e.g., the terms "Patient" and "Examination" in Figure 3 belong to different categories, person and procedure, respectively; (ii) the concept is not yet a part of the SNOMED CT, or there is no linguistic match between the term and the description of the desired concept. In such cases, the category picker module picks the next highest ranked category and repeats until a concept is retrieved, or the top k classes have been explored, where k is an arbitrarily chosen number between 1 and 15.

4. EMPIRICAL EVALUATION

In the previous section, we presented an approach to map a term used in a clinical form to a unique SNOMED CT concept. The approach is based on the hybridization of existing linguistic techniques with the semantic structure of forms. In this section, we present an evaluation of the approach using several clinical forms collected from 5 institutions. The primary goal is to investigate whether employing the semantic structure can improve the mapping performance. Alongside, we also study the kind of impact linguistics make on the mapping performance. We compare the performance of the proposed hybrid approach with that of a baseline linguistic-based approach, and find that the hybrid approach attains a

Table 3: Data Description

Set	Tot. Forms	Total Terms	Mappability(%)
1	3	161	75.77
2	6	261	63.98
3	7	294	56.80
4	5	388	59.02
5	5	397	62.21
All	26	1501	63.55

significant improvement in performance over a baseline linguistic based approach. Also, the hybrid approach achieves a precision of 0.89 and a recall of 0.76. Upon testing multiple combinations of linguistic and structural information, we find that while linguistics only help in improving the recall, structural information helps improve both the recall and the precision. It is thus desirable to devise hybrid approaches that could address both the diversity and the context challenges associated with the term mapping problem.

4.1 Settings

The data used in the experiments are described in Table 3. This includes 26 forms actively used for patient data collection in 5 medical institutions. The forms contain 1501 crude terms(or phrases) supplied by the clinicians. After data collection, we manually identified a SNOMED CT concept, corresponding to each form term. We found that not all the terms were mappable, i.e., relevant with respect to SNOMED CT. This mostly included the terms such as "no scleral icterus" and "chronic back pain," that correspond to the post-coordinated concepts, and the terms such as "follow up with PCP" and "sent to ER," that partially correspond to certain concepts. Therefore, we also report the *mappability* of the forms, i.e., the percentage of the relevant terms found in the forms. Overall, 954 (63.55%) of the terms are mappable.

To conduct the experiments, we developed a JAVA based implementation of the mapping approach illustrated in Figure 5. We redesigned the forms using the interface of the fEHR system and retrieved the respective semantic `form trees`. To train the `sClassifier`, we used cross-validation across the terms of a given dataset. The classification training dataset was manually tagged with appropriate semantic categories. We heuristically chose the value of k, i.e., the number of top classes to be considered for prediction, as 4. As the API module, we used *SnAPI*, a product provided by the Dataline Software Ltd[1]. In terms of the underlying linguistic techniques, *SnAPI* is the programmatic equivalent of the *Snoflake* browser introduced earlier in Section 1.1. We heuristically chose the threshold for the match-weight function adopted by the *SnAPI* as 0.2.

To assess the performance of the approach, we report precision and recall, wherein precision is the number of correctly mapped terms over the total terms mapped by the system, and recall is defined as the number of terms correctly mapped by the system over the total number of relevant terms found in the SNOMED CT services.

4.2 Experiments and Findings

The main goal of the experiments was to study the impact of using structure on the mapping performance. We conducted experiments using 3 methods; with each method we increased the extent of the structural information utilized.

We first devised a **baseline** approach based on pure linguistics. Given a term, this approach uses the *SnAPI* general mapping functionality and maps the term to the most linguistically matching, i.e., the maximum match-weight, SNOMED CT concept. The **baseline** approach resulted into an average precision and an average recall of 0.63 and 0.45, respectively. Next, we conducted experiments using the proposed **hybrid** approach. Using this approach, the precision ranged from 0.65 through 0.89 and the recall ranged from 0.31 through 0.69, for all the datasets. We further enhanced the structure based component of the **hybrid** approach by expanding the candidate set of the semantic categories considered. In particular, we modified the category picker module such that it first retrieves the most linguistically matching concepts for all the top k classes; then, among the candidates, picks the maximum match-weight concept. We call this the **hybrid++** approach. For this method, the mapping precision ranged from 0.78 through 0.92, and the recall ranged from 0.43 through 0.74.

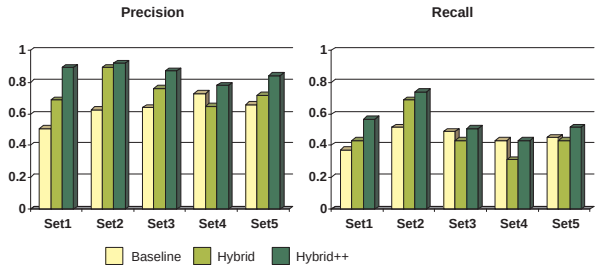

Figure 6: Mapping Performance

Figure 6 summarizes the institution-wise results of the experiments conducted using the 3 methods. The **hybrid** approach improves the precision performance over the **baseline** approach for 4 out of 5 datasets. On an average, the precision improved by 18%. The recall improved by at least 15% for the first two datasets and decreased by 14% for the latter three. The second hybrid approach, **hybrid++**, further improved the performance over the first hybrid approach on an average by 16% in terms of the precision, and by 23% in terms of the recall.

The **hybrid++** method achieved an average precision of 0.86 and an average recall of 0.55. We investigated the reasons for a low recall. It was found that since *SnAPI* uses exact string matching as its underlying linguistic technique, the unsuccessful cases occurred because of string mismatch between the term and the concept descriptions. Hence, we added a term processing component that removes special characters (-,#, /, etc.) and performs acronym expansion using a dictionary of 103 frequently used clinical acronyms such as "T"(temperature), "BTL"(Bilateral Tubal Litigation), "VTE" (venous thromboembolism), etc. We reconducted the experiments for the 3 methods, wherein the form terms were processed before being fed into the API module for concept retrieval. As illustrated in Figure 7, the average performances of the three methods improved consistently. The **hybrid++** method, with the term processing component, achieved an average precision of 0.89 and an average recall of 0.76.

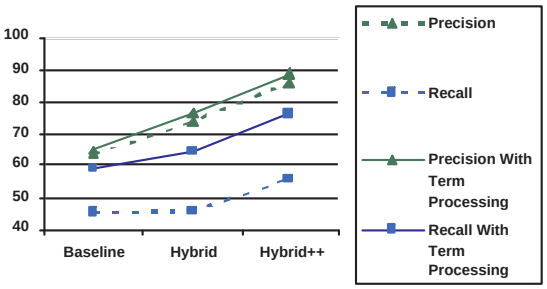

Figure 7: Impact of the term processing component

Finally, we could draw the following implications.

- *Impact of Structure:* The **hybrid** approach involving both structure and linguistics, led to 18% improvement in the precision over the **baseline** approach. On extracting further knowledge from the semantic structure, i.e., using the **hybrid++** method, the average precision improved by 23% and the average recall improved by 38%, over the **baseline** approach. This success is because of the increase in the number of correct concept predictions achieved as a result of incorporating the structural knowledge. This clearly indicates that the structural knowledge has the ability to address the context challenge, and improve the overall mapping performance.

- *Impact of Linguistics:* The impact of linguistics could be quantified by the change in performance upon the addition of the new term mapping component. The new linguistic component improved the precision of the three methods only slightly by 2-3% each. This is depicted by the two close lines for precision in Figure 7. This component, however, had a lot of impact on the recall and improved the performance by at least 30% for all the three methods. This is because the new component helped in retrieving a much larger number of terms. This indicates that linguistic-based approaches can certainly improve the recall and address the diversity challenge to a large extent.

- *Mapping Performance:* Even with the limited training data, the **hybrid++** method with term processing achieved a promising performance with an average precision of 0.89 and an average recall of 0.76. Earlier works have maintained that a high precision can only be achieved using expert analysis [16], yet our automated approach performed well. The relatively lower value of recall could be attributed to the simplicity of the linguistic functions used in this work. The recall improved consistently and significantly upon the addition of the new term processing component. This suggests that it can be further improved by including some external resources such as clinical and general thesaurus, medical acronym dictionary such as RNotes [23], and incorporating some sophisticated term processing techniques such as stemming and auto-correction [25].

5. RELATED WORK

Standardization of clinical data has received a lot of attention in the past. The primary motivation for translating data into standard concepts are to resolve interpretation issues, facilitate clinical and outcomes research, and support future interoperability across systems and institutions. In addition, the research conducted in this area also highlights the usability of the carefully developed medical standards, and implies certain guidelines for refining the standards. In this section, we discuss some of the key works related to mapping freely written clinical data into standards such as SNOMED CT.

The work in Henry et al.[11] investigates whether the narrative description spontaneously entered by nurses, in patient's progress notes and care plans, could be represented by the SNOMED-III terminology. For this investigation, 485 patient encounters are collected from 3 institutions, and the terms describing patient problems are manually extracted. The data to be mapped includes 1841 patient problems composed out of 761 unique terms. The problems are mapped to SNOMED concepts using exact string matching. Overall, 44% of the problems map to a single concept, and 69% map to one or more concepts and allowing the user to choose one. Although the results are not expert validated, it is concluded that it is possible to represent the nursing terms using standard terminology.

Another study [16] performs a mapping between the clinical terms used by the practitioners and an expert designed standard. This study is conducted at the Columbia Presbyterian Medical Center. The data to be mapped consists of the clinical diagnosis and medications information written by practitioners in a clinical profile system. The data terms are either provided by the practitioner, or are chosen from the SNOMED terminology. This data is required to be mapped to a home-grown standard vocabulary, the Medical Entities Dictionary (MED). MED is a semantic network with over 35000 entities, wherein each entity has a name and multiple synonyms. The proposed method creates word groups of the MED entity terms using the lexical variants from the UMLS. Each term from the clinical profile system is tokenized, and matched with the medical entities via the word groups. The possible matches are ranked based on longest common substring similarity with 75% cutoff, and presented to the user. The mapping of 1045 SNOMED-derived terms to entities leads to a recall of 70% and a precision of 61%. Out of the 1225 practitioner supplied terms, 31% map to exactly one entity, and 51% map to at least 1 entity. The results from this dataset are, however, not evaluated.

The work *TokenMatcher*[24] also maps clinical notes containing medical complaints into SNOMED CT concepts. The algorithm pre-processes the notes using sentence boundary detection, term normalization, and POS tagging. A regular expression based entity recognizer is used to extract the relevant terms to be mapped. The algorithm also utilizes an augmented lexicon that consists of a general word to concept mappings derived from the SNOMED CT description table. The key is the token matching step of the algorithm that matches each clinical term to concepts using the augmented lexicon, and assigns a score to each concept description. The algorithm also employs abbreviation expansion using a list of 1254 medical abbreviations. It also performs negation identification by using some rules to identify the post coordinated concepts. The *TokenMatcher* has been deployed as

a web service but no formal evaluation has yet been conducted.

The above mentioned works address the problem of standardizing the clinical notes written for human processing and understanding. In contrast, the work *Model Standardization using Terminology Services,(MoST)*[26] presents a method to map a clinical data model into SNOMED CT concepts. The model considered by *MoST* is the European standard clinical model, known as Archetypes, which is the back bone of the clinical data entry forms. *MoST* is a method to find the candidate SNOMED CT concepts that correspond to the intended meaning of a term used in the data model. The method performs lexical processing of the terms using emergency medical text processing, word sense disambiguation, synonym identification, and term simplification. This is followed by the context processing including identification of the semantic category of the term using UMLS, and mapping the term to concepts using certain filtering rules based on the SNOMED CT categories and relationships. Finally, the modeler is presented with a list of candidate concepts to choose from. The method is tested on 19 models with 475 terms. The precision and recall calculation is relaxed in that any case, where the desired concept is part of the candidate concepts, is considered a success. Overall, the method leads to a recall of 89% and a precision of 82%. After applying the context rules, the precision increases to 90%.

To accomplish mapping, most of the existing works rely on the linguistic similarity techniques such as exploration of synonyms, morphemes, and lexical variants. Such techniques can certainly lead to a large recall. However, the standard vocabularies are growing and getting richer; there are often multiple lexically matching concepts with different semantic intentions, leading to the context challenge introduced in Section 1. It has become increasingly important to accomplish a high precision as well[33].

In this work, we propose that the context-based techniques, when combined with the linguistic techniques, could lead to a higher precision. We propose a method, to map a clinical data entry form to SNOMED CT concepts, which is based on exploiting the semantic structure of forms. Conceptually, our work is similar to *MoST* in that we perform the mapping of clinical meta data as opposed to data. Technically, our work differs as the contextual information used by *MoST* is limited to the SNOMED CT semantic categories. Our work also relies on the context of the form term. Our work is closer to the clinical section classification method proposed by Li et al. [19] that assigns standard labels to the sections of clinical notes by exploiting the organizational structure of the clinical documents. While most of the existing works are semi-automated and only present a candidate list of concepts, our work is completely automated and retrieves a unique concept corresponding to a given form term. In addition, we also conduct a real-world case study on the data-entry forms developed in 5 medical institutions, thoroughly evaluate the results, and draw several insights from the mapping results.

6. CONCLUSIONS

Existing healthcare systems have not been created with future integration in mind. Terms used to name database elements are often not controlled by standard vocabularies. The variety of terms chosen by different users leads to future integration and reconciliation issues. In this paper, we have proposed a structure-based approach to map the user-specified terms, captured in design and data collection artifacts, particularly clinical data-entry forms, into standardized SNOMED CT concepts.

We have introduced and addressed a new problem of mapping a form term to a SNOMED CT concept. While the existing linguistic-based methods are solely based on term-level matching, the proposed method performs a context-level matching followed by a term-level matching. Herein, the context of a given term is systematically extracted from the semantic structure of the form, and the context of a SNOMED CT concept is assumed to be its pre-defined semantic category. The proposed approach first uses a structure-based model to determine the semantic category for a given term, and then maps it to the linguistically matching clinical concept. Another contribution is that we have conducted an empirical study on 26 real-world clinical forms designed by multiple clinicians. Compared to an existing linguistic based approach, the proposed hybrid method achieves a performance improvement of 23% in terms of precision and 38% in terms of recall. The method helps achieve an average precision of 0.89, and an average recall of 0.76. We have studied the individual impact of the structure and the linguistics on the mapping performance. We find that while the term linguistics can only influence the recall performance, the semantic structure has the potential to improve the overall mapping performance. In the future, it is hence desirable to develop hybrid approaches that can address the challenges associated with the mapping problem and lead to a superior performance.

In the future, we intend to quantitatively study the impact of using standard vocabulary-compliant form terms within larger frameworks of form-driven database design [18, 4], in terms of the database annotation extent and the minimization of user interventions. In addition to applying other classification techniques, we intend to address certain limitations of the employed classification model such as handling missing and inapplicable values in the training data. We also intend to study the validity of the assumptions, e.g., that class conditional independence holds true, and that the first concept returned by the category-specific mapping is the desired one. Moreover, we also intend to leverage other defining relationships and the compositional nature of the SNOMED CT, to derive post coordinated mapping expressions, and further improve performance[20]. We also plan to use the other terminologies of the UMLS system[6] and compare their form annotation performances with that of the SNOMED CT.

7. ACKNOWLEDGEMENTS

We sincerely thank the anonymous reviewers for providing valuable suggestions for revising this paper. This research work is supported in part by the grants NSF CCF 0905291, NSF CCF 1049864, and NSFC 90920005.

8. REFERENCES

[1] Dataline software ltd. http://www.dataline.co.uk.

[2] Ihtsdo: International health terminology standards development organisation. http://www.ihtsdo.org/.

[3] SNOMED Clinical Terms User Guide. Technical report, The International Health Terminology Standards Development Organisation, 07 2009.

[4] Y. An, R. Khare, I.-Y. Song, and X. Hu. Automatically mapping and integrating multiple data entry forms into a database. In *In the proceedings of 30th International Conference on Conceptual Modeling*, 2011.

[5] Y. An, J. Mylopoulos, and A. Borgida. Building Semantic Mappings from Databases to Ontologies. In *Proceedings of American Association for Artificial Intelligence (AAAI)*, 2006.

[6] O. Bodenreider. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270, 2004.

[7] R. Elmasri and S. B. Navathe. *Fundamentals of Database Systems, 3rd Ed.* Addison-Wesley, 2000.

[8] S. Gupta, G. E. Kaiser, P. Grimm, M. F. Chiang, and J. Starren. Automating content extraction of html documents. *WORLD WIDE WEB - INTERNET AND INFORMATION SYSTEMS*, pages 179–224, 2005.

[9] A. Halevy. Why Your Data Won't Mix. *ACM Queue*, 3(8):50–58, 2005.

[10] J. Han and M. Kamber. *Data Mining: Concepts and Techniques, 2nd ed.* Morgan Kaufmann, 2006.

[11] S. B. Henry, K. E. Campbell, and W. L. Holzemer. Representation of nursing terms for the description of patient problems using snomed iii. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 700–704, 1993.

[12] M. A. Hernández, L. Popa, H. Ho, and F. Naumann. Clio: A schema mapping tool for information integration. In *Proceedings of ISPAN*, 2005.

[13] S. Hina, E. Atwell, and O. J. and. Secure information extraction from clinical documents using snomed ct gazetteer and natural language processing. In *Proceedins of International Conference for Internet Technology and Secured Transactions (ICITST)*, pages 1–5, 2010.

[14] S. Jean, H. Dehainsala, D. N. Xuan, G. Pierra, L. Bellatreche, and Y. Aït-Ameur. OntoDB: It is Time to Embed your Domain Ontology in your Database. Demo presentation, 2007.

[15] E. Jiménez-ruiz, B. C. Grau, I. Horrocks, R. Berlanga, and R. Berlanga. Logic-based assessment of the compatibility of umls ontology sources. In *Journal of Biomedical Semantics*, pages 1–6, 2011.

[16] R. C. B. Jr., J. J. Cimino, and P. D. Clayton. Mapping clinically useful terminology to a controlled medical vocabulary. In *Proceedings of Annual Symposium of Computing Applications in Medical Care*, pages 211–215, 1994.

[17] R. Khare, Y. An, and I.-Y. Song. Understanding search interfaces: A survey. *SIGMOD Record*, 39(1):33–40, 2010.

[18] R. Khare, Y. An, I.-Y. Song, and X. Hu. Can clinicians create high-quality databases? a study on a flexible electronic health record (fehr) system. In *Proceedings of 1st ACM International Health Informatics Symposium (IHI)*, 2010.

[19] Y. Li, S. Lipsky Gorman, and N. Elhadad. Section classification in clinical notes using supervised hidden markov model. In *Proceedings of the 1st ACM International Health Informatics Symposium*, IHI '10, pages 744–750, New York, NY, USA, 2010. ACM.

[20] R. B. Llavori, E. Jiménez-Ruiz, V. Nebot, and I. Sanz. Faeton: Form analysis and extraction tool for ontology construction. *IJCAT*, pages 224–233, 2010.

[21] J. Madhavan, P. A. Bernstein, and E. Rahm. Generic schema matching with cupid. In *VLDB '01: Proceedings of the 27th International Conference on Very Large Data Bases*, pages 49–58, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[22] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[23] E. Myers. *Rnotes: Nurse's Clinical Pocket Guide*. F. A. Davis Company, second edition, 2006.

[24] J. Patrick, Y. Wang, and P. Budd. An automated system for conversion of clinical notes into snomed clinical terminology. In *In Proc. of HKMD-07, volume 68 of CRPIT*, pages 219–226, 2007.

[25] R. Rada, B. Blum, E. Calhoun, H. Mili, H. Orthner, and S. Singer. A vocabulary for medical informatics. *Comput. Biomed. Res.*, 20:244–263, June 1987.

[26] A. R. Rahil Qamar and. Most: A system to semantically map clinical model data to snomed-ct. In *In the proceedings of Semantic Mining Conference on SNOMED-CT*, pages 38–43, 2006.

[27] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *VLDB JOURNAL*, 10:2001, 2001.

[28] J. Rogers and O. Bodenreider. Snomed ct: Browsing the browsers. In R. Cornet and K. A. Spackman, editors, *Proceedings of Conference in Knowledge Representation in Medicine*, 2008.

[29] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 2003.

[30] H. Stenzhorn, E. J. Pacheco, P. Nohama, and S. Schulz. Automatic mapping of clinical documentation to snomed ct. In *MIE*, pages 228–232, 2009.

[31] V. Sugumaran and V. C. Storey. An ontology-based framework for generating and improving database design. In *Proceedings of the 6th International Conference on Applications of Natural Language to Information Systems-Revised Papers*, NLDB '02, pages 1–12, London, UK, 2002. Springer-Verlag.

[32] A. Tordai, J. van Ossenbruggen, G. Schreiber, and B. J. Wielinga. Aligning large skos-like vocabularies: Two case studies. In *ESWC (1)'10*, pages 198–212, 2010.

[33] L. W. Wright, H. K. G. Nardini, A. R. Aronson, and T. C. Rindflesch. Hierarchical concept indexing of full-text documents. *J. Am. Soc. Inf. Sci*, 50:514–523, 1999.