Semantics Driven Approach for Knowledge Acquisition from EMRs

Sujan Perera*, Cory Henson*, Krishnaprasad Thirunarayan*, Amit Sheth* and Suhas Nair†
*Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis)
Wright State University, Dayton OH, USA
{sujan, cory, tkprasad, amit}@knoesis.org
†ezDI, LLC
suhas.nair@ezdi.us

***Abstract*—Semantic computing technologies have matured to be applicable to many critical domains such as national security, life sciences, and health care. However, the key to their success is the availability of a rich domain knowledge base. The creation and refinement of domain knowledge bases poses difficult challenges. The existing knowledge bases in the health care domain are rich in taxonomic relationships, but they lack non-taxonomic (domain) relationships. In this paper, we describe a semi-automatic technique for enriching existing domain knowledge bases with causal relationships gleaned from Electronic Medical Records (EMR) data. We determine missing causal relationships between domain concepts by validating domain knowledge against EMR data sources and leveraging semantic-based techniques to derive plausible relationships that can rectify knowledge gaps. Our evaluation demonstrates that semantic techniques can be employed to improve the efficiency of knowledge acquisition.**

I. Introduction

Semantic computing technologies have matured to be applicable to many critical domains, such as national security, life sciences, and health care. Semantic technologies are traditionally used to overcome the heterogeneity challenges in data integration, but now it's use is being discussed in the context of finding complex relationships among concepts[1], intelligence applications that are critical to national security[4][5], and answering complex queries by deriving otherwise implicit knowledge. The key to the success of semantic technologies is the availability of rich background knowledge.

A knowledge base consists of domain concepts and their relationships. Although representing domain concepts is important, the relationships that exist between these concepts play a crucial role in realizing the full potential of semantic computing[6]. The relationships are of two types, namely, taxonomic (hierarchical) relationships and non-taxonomic (domain) relationships. While existing knowledge bases in the health care domain[8][7] are rich in hierarchical relationships, they are lacking in domain relationships. Creating a knowledge base with domain relationships requires significant input from domain experts, since automatic methods to extract domain relationships are not accurate as methods developed for extracting hierarchical relationships.

People have used traditional methods such as interviewing domain experts, finding facts from literature, and validating known facts with existing data/use cases to collect the knowledge required to build knowledge bases. These methods are similar to those of requirements elicitation in the software development life cycle and have proven to be inefficient. The knowledge collected by these methods is observed to be subjective, ambiguous, and incomplete. This is particularly true in a domain like health care, since an individual's knowledge significantly depends on his/her experience.

Motivated by the lack of domain relationships in the health care knowledge bases, and considering the opportunities and the challenges associated with building such knowledge bases, we proposed a data driven methodology to collect relevant domain knowledge[9]. The availability of domain experts in the health care domain is a rare asset, hence it is necessary to use their availability efficiently. For example, imagine that we have identified 50 disorders and 100 symptoms, and our task is to identify the causal relationships between them. There are 5,000 $(50 * 100)$ possible relationships in this scenario, and the domain expert is forced to perform the tedious task of going through each of them to determine the valid ones. Instead, our method identifies likely disorders that can account for a symptom by analyzing real world data and enrich the background knowledge[9]. It reduces the burden on domain experts by identifying plausible relationship instances that need validation. We will use the term '*data driven method*' to refer this method from here onwards.

The *data driven method*[9] assumed that each symptom in an EMR document should be explained by at least one disorder present in the document. This can happen if there is a causal relationship between the explanatory disorder and the symptom in the knowledge base. If there is no such disorder present in the document, we identify this symptom as an "unexplained symptom", and the *data driven method* predicts all co-occurring disorders as candidates to have a causal relationship with the symptom. We extend this work by proposing an algorithm to select a subset of disorders from a co-occurring disorder set that are most likely to have a causal relationship with the symptom by leveraging knowledge base. Specifically, we make use of both explicitly given causal relationships between symptoms and disorders, and hierarchical relationships among the disorders. The intuition behind our approach is that similar disorders manifest similar symptoms. Our approach extracts disorders that are known to have a causal relationship with the symptom from the knowledge base and collects "similar" disorders by exploiting hierarchical relationships. Then, it checks for an overlap between the disorders that co-occur with the unexplained symptom in an EMR and the collected "similar" disorders to obtain the most plausible disorders that can explain the symptom. Our evaluations demonstrate that this method improves the precision

of the suggested relationships significantly and holds promise for good recall given more data. This ultimately improves the efficiency of the knowledge acquisition activity from domain experts since we now need to ask far less questions to acquire the required knowledge.

The contributions of this work include:

1) A method to validate the richness of a knowledge base relative to a given data set,
2) A method to detect the absence of causal relationships in a given knowledge base, and
3) An efficient, semi-supervised method to suggest new relationships that can rectify the missing relationships.

II. Related Work

Researchers have used different techniques to identify non-taxonomic relationships from the existing knowledge bases and from the literature.

A comprehensive ontology evaluation framework proposed in[14] uses Scarlet[13] to find the relationships among the concepts. Scarlet[13] uses multiple rules to derive taxonomic relationships as well as domain specific relationships. While these rules are capable of deriving taxonomic relationships by integrating multiple ontologies, it cannot derive non-taxonomic relationships. Scarlet can find such a relationship only if some other existing ontology expresses such knowledge.

People use freely available knowledge(peer reviewed publications, books, articles, etc.) to glean the domain ontologies. The most popular techniques to learn the ontologies from the text corpus are based on the Natural Language Processing(NLP) and Machine Learning(ML). NLP and ML based techniques are used to identify the domain entities and taxonomic and non-taxonomic relationships. These techniques rely on named entity identification methods[17], predefined linguistic patterns[17][19][20], lexical syntactic properties of the free text (like frequency of words appearing together[17][16][20], position of the words[15], and frequency of verbs appearing with the lexical terms[16][18]).

Causality is an important relationship for multiple reasons and has gained much attention in contemporary literature. People have developed techniques to mine the causal relationships from text and the majority of these techniques follow the same methods mentioned above. [22][23][24] used a syntactic patterns and [21] used a co-occurrence based method to identify causal relationships. [25] proposes a method to identify causal relationships by using part-of relationships. It claims that a causal relationship can be identified by using fine-grained events.

Our work is different from above methods as it can detect the **absence** of causal relationship in the knowledge base and **suggest** plausible relationships that can rectify absent relationships. This capability can complement existing methods by providing the candidate relationships to look for in the literature/knowledge bases. This will help existing algorithms to have a better focus on domain knowledge exploration.

III. The Approach

We propose an efficient approach based on the IntellegO ontology[10] to identify and acquire missing causal relationships between disorders and symptoms. We use the IntellegO ontology and EMR documents to assess the richness of the knowledge base, find the symptoms which lack relevant relationships, and suggest suitable relationships to rectify those gaps. The suggestions are formulated as "yes/no" questions for domain experts to answer as follows.

Is 'symptom A' a symptom of 'disorder B'?

Our goal is to minimize the number of questions posed and maximize the acquisition of missing relationship instances. Hence the questions we generate should correspond to highly plausible relationships between disorders and symptoms(i.e., we should minimize the number of questions answered with "no"), to use the domain expert's time efficiently. We briefly discuss the application of the IntellegO ontology in the following section.

A. Ontology of Perception (IntellegO)

Perception is the process of interpreting observations of the environment to derive situational awareness; or, in other words, the process of translating low-level observations into high-level abstractions. IntellegO is an ontology that provides formal semantics of machine perception by defining the informational processes involved in translating observations into abstractions. The ontology is encoded in set-theory and has been used in various applications[10][11].

Diagnosis is a function of perception. Medical professionals derive disorders (abstractions) by examining symptoms (low level signals). EMR documents implicitly contain the knowledge involved in the this informational process. This knowledge is implicit in that EMR documents do not mention that diagnosis A has been derived by observing symptoms B, C and D, but they just mention the concepts A, B, C and D. We used IntellegO ontology because it nicely aligns with the characteristics of the knowledge that we want to represent in the health care domain, improves interoperability, and supports "perceptual" reasoning.

The proposed algorithm uses a subset of concepts (*'intellego:entity'*, *'intellego:quality'*, *'intellego:percept'* and *'intellego:explanation'*) and the *'intellego:inheres-in'* relationship from the IntellegO ontology[1]. Furthermore, it introduces the new concept *'intellego:coverage'*. The semantics of *'intellego:coverage'* is defined in Section III-B4. Here we discuss the semantics of the above mentioned concepts.

Let us take *hypertension* and an associated symptom, *chest pain*. *hypertension* is an *'intellego:entity'* and *chest pain* is an *'intellego:quality'*. The *chest pain 'intellego:quality'* is an inherent property of *hypertension 'intellego:entity'*. In general, *'intellego:entity'* is an object or event in the real world and *'intellego:quality'* is an inherent property of an *'intellego:entity'*. *'intellego:inheres-in'* is the relationship between *'intellego:quality'* and *'intellego:entity'*.

The perception process begins by observing a few qualities. From these observations, it derives entities which can explain all observed qualities. *chest pain* can be explained by the presence of *hypertension*[2]. The set of observed *'in-*

[1] intellego prefix specifies terms from the IntellegO ontology

[2] Note that there may be multiple entities that can explain the observed qualities (e.g., *chest pain* can be explained by *hypertension*, *cardiomyopathy*, *coronary artery disease* and a host of other disorders), but for simplicity, we assume *chest pain* can be explained only by *hypertension*.

tellego:quality' (i.e., *chest pain*) are members of the class *'intellego:percept'* and *'intellego:entity'* (e.g., *hypertension*) which can explain the *'intellego:percept'* are members of the class *'intellego:explanation'*. *'intellego:explanation'* and *'intellego:percept'* are sub-classes of *'intellego:entity'* and *'intellego:quality'* respectively.

As illustrated in this section, the concepts from the IntellegO ontology map to the concepts in the health care domain as[3]; *'intellego:entity'* to `DISORDER`, *'intellego:quality'* to `SYMPTOM`, *'intellego:percept'* to `OBSERVED_SYMPTOM`, *'intellego:explanation'* to `EXPLANATORY_DISORDER`, and *'intellego:inheres-in'* to `IS_SYMPTOM_OF` relationship. We will use these mapped terms instead of IntellegO classes to improve readability.

B. Gleaning Plausible but Missing Causal Relationships

We have developed a system to detect the absence of causal relationships in existing knowledge bases, and suggest the most likely relationships in the form of questions, to rectify these missing relationships. We used real world data sources(i.e., EMR documents) to generate the questions. An EMR document is *consistent* if the symptoms appearing in the document are accounted for by the disorders in it, otherwise it is *inconsistent*. We show how to determine whether a document is consistent or not in Section III-B5. Throughout this process, we assume that the EMR documents are consistent. This is not always true, but the "signal" is strong enough to deal with the "noise".

Our approach consists of the following steps and they are discussed in detail later.

1) Build the initial knowledge base.
2) Semantically annotate the EMR documents with concepts from the knowledge base.
3) Populate `OBSERVED_SYMPTOM` and `EXPLANATORY_DISORDER` for each document.
4) Generate *'intellego:coverage'* for each document.
5) Identify inconsistent EMR documents[4].
6) Suggest plausible candidate relationships between disorders and symptoms synthesized from the EMR documents that can rectify the inconsistencies.
7) Validate the suggested relationships by consulting a domain expert.
8) Update the knowledge base based on expert feedback.
9) Repeat steps 6, 7 and 8 until no new question is generated or satisfied with current results.

1) Build the Initial Knowledge Base

We built an initial knowledge base with minimum involvement from domain experts. We selected a set of concepts to be included in knowledge base based on the frequency of their appearance in the EMR corpus. Then we used the UMLS semantic types to categorize these concepts into symptoms and disorders. The concepts belonging to semantic types "Finding (T033)" and "Sign or Symptom (T184)" are categorized as symptoms and concepts belonging to "Disease and Syndrome (T047)" are categorized as disorders. There were few concepts that do not belong to these categories. We consulted a domain expert to categorize the semantic types of such concepts into disorder or symptom. For example, *atrial fibrillation* belongs to "Pathologic Function (T046)" in UMLS and is categorized as disorder by the domain expert. Our initial knowledge base consists of 86 disorders and 42 symptoms. The next task is to identify an initial set of domain relationships. We were able to identify 255 relationships between these symptoms and disorders with the help of domain experts.

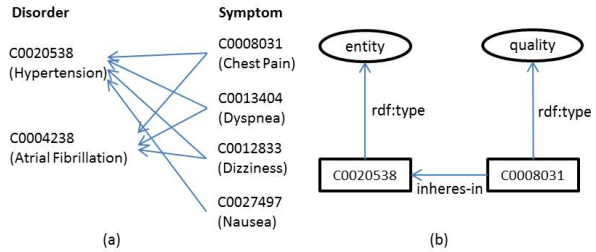

Fig. 1. inheres-in relationships between disorders and symptoms and their representation in IntellegO

The identified disorders were added as individuals of `DISORDER`, and symptoms as individuals of `SYMPTOM`. The `IS_SYMPTOM_OF` relationship is used to represent the causal relationship between disorders and symptoms. We used the Concept Unique Identifier (CUI) to represent the concepts. CUIs are defined in UMLS and widely used in the health care domain. Our EMR documents are annotated using the UMLS vocabulary, as described in the next section, Usage of CUIs allows us to have direct mapping between EMR documents and the knowledge base. Furthermore, this helps to resolve the heterogeneity problem that arises in data representation. For example, EMR documents may use different terms to denote the same concept, one document may use the term *shortness of breath*, while another document may use the term *dyspnea* to refer to the same condition. The use of CUIs allows us to normalize these terms since UMLS will have one CUI to represent multiple syntactic forms of same concept(in this case C0013404 for both *dyspnea* and *shortness of breath*). Figure 1(a) shows the nature of the knowledge (this figure shows both forms of the knowledge base to improve readability). Figure 1(b) shows how this knowledge is represented using IntellegO.

2) Semantically Annotate EMR Documents

Semantic annotation is the process of identifying the entities present in the document with a controlled vocabulary. Annotation is used to make unstructured data machine understandable; i.e., make semantics explicit. EMR documents can have disorders, symptoms, medications, and procedures as entities and contain negations and patient's history information extensively. It is important to identify these aspects associated with each concept since we are interested in the conditions that the patient currently (as opposed to historically) has (as opposed to does not have). There are multiple NLP engines which are capable of processing clinical documents. We compared three such engines for our experiment and decided to use cTAKES[12] and MedLEE[2] since MetaMap[3] does not identify the temporal aspect of identified entities.

Both cTAKES and MedLEE output an XML document and

[3]We use the uppercase term to distinguish between domain entity and the corresponding IntellegO class

[4]Whenever we say inconsistent EMR, we mean EMR document is inconsistent w.r.t knowledge base, because the latter cannot explain all the symptoms using the disorders in the document

<condition value="atrial fibrillation" code="49436004:SNOMED" uncertainty="0" polarity="0" conditional="false" cui="C0004238" tui="T046"/>

Fig. 2. xml element describes *Atrial Fibrillation* in cTAKES output

Figure 2 shows an example element from an XML document generated by cTAKES. The annotations of Figure 2 are interpreted for the disorder *atrial fibrillation* as indicating the presence of disorder(polarity="0") right now(uncertainty="0"). cTAKES reports 0.80 and MedLEE reports 0.89 for entity identification task and the recall values are 0.64 and 0.84 for cTAKES and MedLEE respectively. The detailed results of the evaluation is reported in [12] and [26] for cTAKES and MedLEE respectively.

3) Populate `OBSERVED_SYMPTOM` *and* `EXPLANATORY_DISORDER`

Semantic annotation allows us to populate `OBSERVED_SYMPTOM` and `EXPLANATORY_DISORDER`. Both NLP engines do not distinguish between symptoms and disorders; it generates one type of XML element as shown in Figure 2 for both symptom and disease. We use our initial knowledge base to distinguish symptoms and disorders in the annotated documents. The EMR document has multiple sections such as 'current diagnosis', 'review of systems', 'physical examination', 'history of current illness', 'family history', and 'assessment and recommendation'. We use symptoms and disorders mentioned only in the 'current diagnosis', 'review of systems', 'physical examination', and 'assessment and recommendation' sections, because these sections contain information most relevant to the patient's current status. Each 'condition' node in an XML document which is of type `SYMPTOM` is annotated as `OBSERVED_SYMPTOM` and each 'condition' node which is of type `DISORDER` is annotated as `EXPLANATORY_DISORDER`. Recall that `SYMPTOM` present in an EMR document is an `OBSERVED_SYMPTOM` and `DISORDER` present in EMR document should be able to explain the set of `OBSERVED_SYMPTOM`s, hence they belong to `EXPLANATORY_DISORDER`.

4) Generate 'intellego:coverage'

Coverage can be defined as the aggregation of `SYMPTOM`s that can be accounted for by a set of `EXPLANATORY_DISORDER`.

Formally, coverage is defined as, $intellego : coverage \equiv \exists \texttt{IS\_SYMPTOM\_OF}\{e_1\} \sqcup \exists \texttt{IS\_SYMPTOM\_OF}\{e_2\} \sqcup \cdots \sqcup \exists \texttt{IS\_SYMPTOM\_OF}\{e_n\}$

where $e_i, i = 1, 2, 3, ..., n$ are instances of `EXPLANATORY_DISORDER`.

The following example defines a coverage class for an EMR document which reports *atrial fibrillation* and *hypertension* as disorders(we restrain from using CUIs to improve readability).

$intellego : coverage \equiv \exists \texttt{IS\_SYMPTOM\_OF}\{hypertension\} \sqcup \exists \texttt{IS\_SYMPTOM\_OF}\{atrial fibrillation\}$

We use OWL reasoner to populate the instances of *'intellego:coverage'* class. We will use the term `COVERED_SYMPTOM` to denote the class *'intellego:coverage'* from here onwards.

5) Identify inconsistent EMR documents

As stated earlier, we assume that the EMR documents are consistent, i.e., the symptoms appearing in the document are accounted for by the disorders in it. Formally, the *isConsistent* function is defined as,

$$isConsistent = \begin{cases} true & \text{if } \texttt{OBSERVED\_SYMPTOM} \subseteq \texttt{COVERED\_SYMPTOM} \\ false & \text{otherwise} \end{cases}$$

We have identified the following factors that may cause an inconsistency:

- The text conversion can introduce errors.
  The input to our algorithm is structured(XML) document which is generated by NLP engines. The conversion of text documents to XML documents may introduce errors. E.g., the phrase 'The patient's symptoms of shortness of breath and chest discomfort has resolved' can incorrectly result in an XML element that indicates a presence of *shortness of breath* and *chest discomfort*. This causes unexpected symptoms to appear in the XML document.
- Some combination of disorders can produce symptoms that are not inherent in any individual disorders. The knowledge base does not represent complex situations where multiple disorders can manifest new symptoms(over and above those caused by individual disorders) through complex interactions over time. Presence of such scenarios can include extra symptoms in `OBSERVED_SYMPTOM` that cannot be in `COVERED_SYMPTOM`.
- Irrelevant observations,
  An EMR document can contain symptoms that are not used for diagnosis, causing inconsistency.
- Missing domain knowledge (i.e., missing causal relationships between `SYMPTOM` and `DISORDER`).
  The accuracy and completeness of `COVERED_SYMPTOM` depends on the accuracy and completeness of the knowledge base. If the knowledge base lacks a relationship, then the generated `COVERED_SYMPTOM` set can be incomplete.
  E.g., Assume that a *hypertension* patient has *edema*, but the relationship between *hypertension* and *edema* (*edema* `IS_SYMPTOM_OF` *hypertension*) is not present in the knowledge base. This leads to `OBSERVED_SYMPTOM` not being a subset of `COVERED_SYMPTOM` in the patient EMR document.

Our proposed method uses missing domain knowledge as an entry point to generate questions which ultimately finds and validates (with the help of domain expert) new knowledge to remedy missing relationships.

6) Suggest Candidate Relationships

Identification of inconsistent EMR documents leads us to detect the absence of causal relationships and generate plausible candidate relationships between symptoms and disorders which should be validated by a domain expert. When we find a symptom that appears in `OBSERVED_SYMPTOM` but not in `COVERED_SYMPTOM`, all disorders in `EXPLANATORY_DISORDER` become candidates to be related to an unexplained symptom. We call this set 'candidate disorder set'. We use prior knowledge about the symptom and the hierarchical relationships of the disorders to filter out the most plausible disorders from the 'candidate disorder set' to have causal relationship with the symptom.

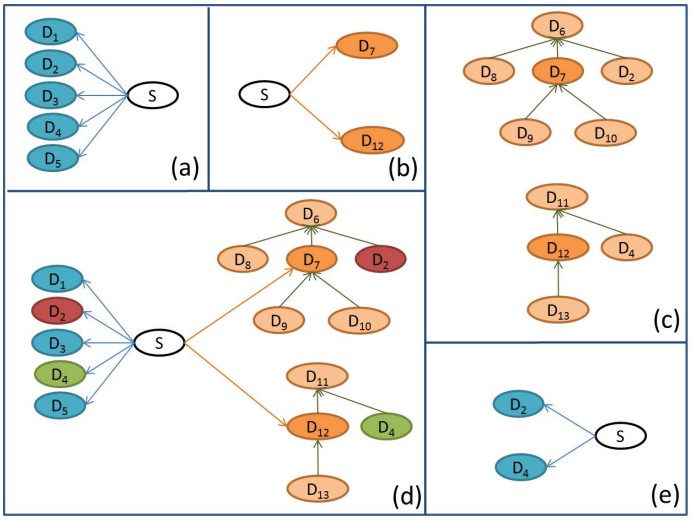

Fig. 3. Suggest Candidate Relationships

The following steps describe how we suggest candidate relationships.

1) Collect a set of disorders from our initial knowledge base which are related to the unexplained symptom. We will refer this disorder set as 'known disorders' in what follows.
2) Collect disorders that appear in the "neighborhood" of each known disorder. We use the UMLS concept hierarchy to collect the neighborhood of a particular disorder. Specifically, we collect parents, children, and siblings of a disorder as its neighborhood.
3) Union the collected neighborhoods and intersect that with the disorders in the EMR documents to obtain the filtered candidate disorder set.

Figure 3 depicts the steps in this process. Let symptom S be unexplained in the document, and disorders D_1, D_2, D_3, D_4, D_5 are present in the same document (Figure 3(a)). Initially all these disorders are members of the candidate disorder set. From the initial knowledge base, we find that symptom S is a symptom of two other disorders, namely D_7 and D_{12}(Figure 3(b)). With this extra knowledge, we collect the neighborhoods for D_7 and D_{12} by using the UMLS concept hierarchy as depicted in Figure 3(c). It turns out that D_2 and D_4 are members of both initial candidate set and collected neighborhoods (Figure 3(d)). This suggests that D_2 and D_4 are the most probable candidates that can explain symptom S, and results in eliminating D_1, D_3 and D_5 from the initial candidate set (Figure 3(e)).

The intuition behind this method is: a symptom is most likely to be shared by similar disorders. We collect similar types of disorders by exploiting the UMLS concept hierarchy.

Selected candidate relationships are presented to the domain expert in the form of questions. Then the knowledge base is updated according to the feedback from the domain expert, i.e., if the expert agrees with suggested causal relationship between the disorder and the symptom, it is added to the background knowledge, otherwise, it is ignored. Since this step adds more knowledge to the knowledge base about the unexplained symptom, we run the 'Suggest Candidate Relationships' as many times as necessary. The algorithm terminates when there are no new questions generated. Since there are only finite number of elements in the candidate set, the termination of the algorithm is guaranteed.

IV. Experiment

We implemented the proposed algorithm in Java, and the OWL API[5] was used to interact with the ontology. The Pellet reasoner[6] was used for the reasoning task of finding the instances of class `COVERED_SYMPTOM`. We conducted the evaluation using a set of de-identified EMR documents[7]. The corpus consists of 1500 unstructured EMR documents.

We parsed these 1500 EMR documents using MedLEE and cTAKES and executed our algorithm on both parsed corpora separately. Here we discuss and compare the results of our algorithm on the two corpora.

We executed our algorithm on each parsed document and collected corpus wide statistics. Specifically, we determined within how many documents a particular symptom was found as unexplained and what disorders co-occurred with the symptom in such instances and their frequency. Whenever we find symptom S unexplained, we call it an *unexplained instance* of symptom S. We found 23 unexplained symptoms from MedLEE corpus and 29 unexplained symptoms from cTAKES corpus. Each of these unexplained symptoms co-occurs with multiple diseases with varying frequency within the corpus. Table I contains the top 5 unexplained symptoms of both parsed corpora. Table II contains top co-occurring disorders for *edema* when *edema* was found as unexplained, in each corpus. According to Table I, *edema* is found as unexplained in 206 documents parsed with MedLEE and 910 documents parsed with cTAKES. Hence there are 206 unexplained instances of *edema* in the MedLEE corpus and 910 such instances in the cTAKES corpus. The first row of Table II says that *hyperlipedemia* was present 116 times within those 206 instances in the MedLEE corpus and *hypertension* was present 647 times within those 910 instances in the cTAKES corpus.[8]

Once we identified the unexplained instances of symptoms and co-occurring disorders in each instance as above, the next task is to suggest the relationships that can rectify the unexplained instances. We evaluated the efficiency of the relationships suggestion capability of our algorithm in following three dimensions.

1) The Precision of the Suggested Relationships.
2) The Recall of the Suggested Relationships.
3) The Increment of the Explanatory Power of the Knowledge Base.

A. The Precision of Suggested Relationships

One of our main goals is to make effective and efficient use of the domain expert's availability for knowledge acquisition. Hence we would like to maximize the likelihood of suggesting "valid" relationships. In other words we would like to improve

[5]http://owlapi.sourceforge.net/
[6]http://clarkparsia.com/pellet/
[7]The IRB allows us to use the data in this study, but does not allow us to release it for public use yet.
[8]Extended results can be found at http://knoesis.org/researchers/sujan/experiments.html

MedLEE		cTAKES	
Symptom	# of unexplained	Symptom	# of unexplained
edema	206	edema	910
depression	172	syncope	336
angina	134	systolic murmur	168
dyspnea	120	tachycardia	143
syncope	103	angina	136

TABLE I
TOP 5 UNEXPLAINED SYMPTOMS

MedLEE		cTAKES	
Disorder	# of times	Disorder	# of times
hyperlipidemia	116	hypertension	647
hypertension	112	hyperlipidemia	641
atrial fibrillation	84	claudication	454
coronary artery disease	66	coronary atherosclerosis	395
coronary arteriosclerosis	62	coronary artery disease	242

TABLE II
DISTRIBUTION OF DISORDERS CO-OCCURRING WITH UNEXPLAINED *edema*

the precision of the suggested relationships. The *precision* for our experiment is defined as,

$$precision = \frac{number\text{-}of\text{-}suggested\text{-}correct\text{-}links}{total\text{-}number\text{-}of\text{-}suggested\text{-}links} * 100$$

Table III summarizes the precision of suggested relationships for each corpus. The suggested relationships in the first round with the MedLEE corpus has a precision of 77.55%. The precision of the cTAKES corpus is 73.94%. Iteration 2 added newly found relationships in the first round to two different knowledge bases forked from the initial knowledge base. The first copy is enriched with 76 relationships found from the MedLEE corpus and the second copy is enriched with 105 new relationships found from the cTAKES corpus. The algorithm was executed with these new knowledge bases. This iteration suggested 16 new relationships for the MedLEE corpus and 29 new relationships for the cTAKES corpus.

We decided to terminate the algorithm after the 2nd iteration since the 3rd iteration added only 4 relationships with poor precision. The overall experiment (two iterations) suggested a total of 114 relationships for the MedLEE corpus out of which 86 were correct, yielding a precision of 75.43%. The experiment suggested 171 relationships for the cTAKES corpus out of which 125 were correct, yielding a precision of 73.09%.

A much simpler method to find these relationships is to suggest all co-occurring diseases when a symptom is found to be unexplained. The strength of the proposed method over this simpler method is its ability to filter out disorders which do not have a causal relationship with the symptoms, even though the disorders co-occur with the unexplained symptom. Here we demonstrate this by comparing the results obtained

Iteration	corpus	# of suggestions	# of correct	*precision*
1	MedLEE	98	76	77.55%
	cTAKES	142	105	73.94%
2	MedLEE	16	10	62.5%
	cTAKES	29	20	68.96%
3	MedLEE	8	4	50.0%
	cTAKES	9	4	44.44%

TABLE III
THE PRECISION OF SUGGESTED RELATIONSHIPS

Symptom	# of co-occurring disorders(COD)	# of correct in COD	# of incorrect in COD	# of correct suggestions	# of incorrect suggestions
angina	7	5	2	4	0
chest pain	3	0	3	0	0
nausea	3	1	2	1	0
fatigue	3	2	1	2	1
dyspnea	3	0	3	0	1
...	...	...	...	...	...
Total	**103**	**37**	**66**	**25**	**7**

TABLE IV
COMPARISON OF THE MEDLEE OUTPUT WITH A SIMPLER METHOD

Symptom	# of co-occurring disorders(COD)	# of correct in COD	# of incorrect in COD	# of correct suggestions	# of incorrect suggestions
chest pain	3	0	3	0	0
numbness	8	3	5	1	1
nausea	5	1	4	1	1
dyspnea	6	1	5	0	2
angina	11	5	6	4	1
...	...	...	...	...	...
Total	**200**	**58**	**142**	**31**	**15**

TABLE V
COMPARISON OF THE CTAKES OUTPUT WITH A SIMPLER METHOD

by such a simpler method with results obtained by our method. As mentioned before, we found 23 unexplained symptoms in the MedLEE corpus and 29 unexplained symptoms in the cTAKES corpus. Each of these unexplained symptoms co-occur with more than one disorder. The 23 symptoms have a total of 465 co-occurrences with different disorders in the MedLEE corpus and 29 symptoms have a total of 947 such co-occurrences in the cTAKES corpus. But due to the limited availability of the domain experts, we were not able to validate all these co-occurrences. Hence we decided to validate the top co-occurring disorders (based on co-occurrence frequency) of each symptom and compare the results with our method. Specifically, we collected the top 20% of the co-occurring disorders of each symptom and asked domain experts to mark the correct causal relationships among them. Then we calculated how many of them are found by our method. Table IV and Table V show the results of this experiment for 5 symptoms[9]. According to Table IV, the top 20% of co-occurring disorders with unexplained *angina* consist of 7 disorders. Within these 7 disorders, 5 of them have causal relationships with *angina*. Our method suggested 4 out of 5 and did not suggest any incorrect relationship.

In summary, there were a total of 103 relationships within the top 20% of co-occurring disorders for each unexplained symptom in the corpus parsed with the MedLEE, out of which 37 were correct. Hence the simpler method would have a precision of 35.92% (37/103). Our method suggests 25 out of 37 correct relationships while suggesting only 7 incorrect relationships with the precision of 78.12% (25/32) and a recall of 67.56% (25/37). The precision value of the simpler method for the corpus parsed with the cTAKES is 29.0% (58/200), while the precision and recall for our method for the same corpus is 67.39% (31/46) and 53.44% respectively. This experiment shows that our method was able to filter out incorrect relationships to improve the precision significantly

[9]Due to the space limitation we do not present results for 23 symptoms of MedLEE and 29 symptoms of cTAKES. The complete results can be found at http://knoesis.org/researchers/sujan/experiments.html

	MedLEE	cTAKES
All Correct Relationships	94	109
Known Correct Relationships	41	43
Found Correct Relationships	20	30
Not Found Correct Relationships	33	37
Recall	37.73%	45.45%

TABLE VI
RECALL OF SUGGESTED RELATIONSHIPS FOR 30 EMRS

	MedLEE	cTAKES
All Correct Relationships	94	109
Known Correct Relationships	41	43
Found Correct Relationships	32	44
Not Found Correct Relationships	21	22
Recall	60.37%	66.67%

TABLE VII
RECALL FOR THE RELATIONSHIPS FOUND IN 30 EMRS WHEN EXECUTED WITH MORE DATA

compared to the simpler method while maintaining good recall.

Each of the suggested relationships in the above experiments and in the experiments that follow were validated by two domain experts. We used the following reliable online resources to resolve domain experts disagreements.

- NLM(www.nlm.nih.gov)
- PubMed
- WebMD(www.webmd.com)
- Cleveland Clinic(www.clevelandclinic.org)
- Wikipedia(www.en.wikipedia.org)
- Mayo Clinic(www.mayoclinic.com)
- Healthline(http://www.healthline.com)

We observed two main reasons for disagreement. 1.) If disease A causes Symptom B and Symptom B causes Symptom C, one domain expert interprets that C is related to only B while other domain expert interprets that C is related to both A and B. 2.) One domain expert states that symptom S is highly related to disorder D while other domain expert states that they are not highly related. Furthermore, they consider this relatedness measure when validating the suggested relationships. The difference in threshold values they use to interpret the status of the relationship leads to disagreements.

B. The Recall of Suggested Relationships

The proposed approach is capable of finding causal relationships between symptoms and disorders that are missing in the given knowledge base but **present in the real EMRs**. Due to the limited availability of the domain experts we could not conduct an experiment to calculate the recall for 1,500 EMRs. Instead, we randomly selected 30 EMR documents and asked domain experts to find all the causal relationships that exist in these documents. For example, if an EMR document contains 3 disorders and 4 symptoms, there are 12 possible relationships. We asked domain experts to select the correct causal relationships among these 12 relationships. Let us say the domain experts found 7 causal relationships, 3 of which are already present in the given knowledge base. Then, we expect our method to find the remaining 4 relationships. Hence we define recall as follows and Table VI shows the results of this experiment.

$$recall = \frac{correct\ relationships\ found}{all\ correct\ relationships\ -\ known\ correct\ relationships} * 100$$

'correct relationship found' is the number of correct relationships found by our method, 'all correct relationships' is the number of all correct relationships among symptoms and disorders and 'known correct relationships' is the number of already known relationships among 'all correct relationships'. In other words, the denominator is the number of unknown correct relationships that exist ("knowledge gaps"), while the numerator is what is uncovered by our method.

We identified two main reasons for low recall in Table VI.

- If at least one disorder explains a symptom in the EMR then our method will miss other possibilities:
  A symptom is not identified as unexplained if there is at least one disorder in the document that can explain the symptom and the knowledge base has this relationship. This prevents suggesting other co-occurring disorders within the document as candidates for causal relationship even if they are causally related to this symptom. For example, consider an EMR document with *edema*, *congestive heart failure*, *hypertension*, and *cardiomyopathies*. The knowledge base has a relationship between symptom *edema* and disorder *hypertension*. This makes *edema* explainable within this document, hence our approach does not suggest the other two disorders as candidates that have a causal relationship with *edema* although they actually have such a relationship. So our approach misses these two causal relationships within this EMR document. But, given more data that do not contain these comorbidities (concurrent disorders) our approach is capable of finding these relationships. For example, if there is an EMR document within the given corpus, which has *edema*, *congestive heart failure*, and *cardiomyopathies*, but not *hypertension*, *edema* becomes unexplained and both *congestive heart failure* and *cardiomyopathies* become plausible candidates for a causal relationship with *edema* and these two relationships can be suggested and eventually added to the knowledge base.
- The neighborhood method cannot reach the disorder:
  Even though the disorder *cardiomyopathies* co-occurs with the symptom *edema* in the above scenario, if none of the neighbors of *cardiomyopathies* have a relationship to *edema* in the current knowledge base, our algorithm cannot reach *cardiomyopathies* in the neighborhood collection step. As a consequence, it will never suggest this relationship.

The experiment with 30 EMR documents missed 33 and 37 correct causal relationships from the MedLEE and the cTAKES corpora respectively as shown in Table VI. The lack of different combination of comorbidities within 30 documents significantly contributes towards this result. Hence, given more data, our method should be able to find these relationships. To demonstrate this, we selected 400 documents from each corpus and executed our algorithm. This time our intention was to check how many of the missed relationships are suggested with more data. Table VII contains the improved recall for relationships found in above 30 documents given more data.

C. The Increment of Explanatory Power of the Knowledge Base

The overall goal of our algorithm is to find the missing causal relationships of the knowledge base and enrich it in a semi-supervised manner. Hence it is necessary to quantify the new knowledge obtained by our approach. Since there is no standard technique to quantify the completeness/richness

	MedLEE		cTAKES	
Knowledge base	UI	increment of EP	UI	increment of EP
initial knowledge base	1314	0%	2251	0%
after iteration 1	820	37.59%	878	60.99%
after iteration 2	790	39.87%	806	64.19%

TABLE VIII
COMPARISON OF EXPLANATORY POWER OF KNOWLEDGE BASES

Original Sentence	MedLEE Interpretation	cTAKES interpretation
no deformities, clubbing, cyanosis, erythema or edema observed	edema:negative	edema:positive
S1 normal, S2 normal, S3 present, S4 present, grade 2/6 systolic murmur	systolic murmur:negative	systolic murmur:positive
The muscle soreness she describes could likely be due to side effects from statin therapy	muscle soreness:positive	soreness:positive

TABLE IX
DIFFERENT INTERPRETATIONS BY NLP ENGINES FOR THE SAME PHRASE

of the knowledge base, we use the explanatory power of the knowledge base as a metric to quantify the completeness. We define the explanatory power of the knowledge base relative to a given dataset as,

EP = # of instances in data set explainable by knowledge base

where EP is Explanatory Power and the increment of explanatory power of a new knowledge base relative to initial knowledge base defined as:

$$\textit{increment of EP} = \frac{UI_i - UI_n}{UI_i} * 100$$

where UI_i is the number of unexplained instances of a given data set with the initial knowledge base and UI_n is the number of unexplained instances with the new knowledge base.

As Table VIII shows(UI stands for number of unexplained instances), the explanatory power of the initial knowledge base is increased by 39.87% at the end of the second iteration with the MedLEE corpus, and it is increased by 64.19% at the end of the second iteration with the cTAKES corpus.

D. A Discussion on Dissimilarities in Results of Two NLP Engines

Throughout the experiment, it is evident that the two NLP engines produce different values for each metric even though we used the same EMR corpus. This section informally discusses the observed reasons for these differences. A formal comparison of the two NLP engines is out of scope of the present work.

There are instances where the semantic interpretations of the same phrase/sentence by two NLP engines do not agree, resulting in the differences for two corpora. Table IX contains illustrative examples, which we will further explain.

1) Dissimilarity of unexplained instance count for each symptom (Refer Table I)

The number of unexplained instances for each symptom in the cTAKES corpus is higher than the MedLEE corpus (Table I). One of the reasons for this is the differences in identification of negation in the two engines as shown in the first two rows of Table IX. This interpretation leads the cTAKES corpus to contain more unexplained *edema* instances than the MedLEE corpus. The high frequency of this sentence pattern within the corpus causes a significant difference in the number of unexplained counts for symptoms like *edema* and *syncope*. The same reason leads to symptoms like *systolic murmur* to be present in the cTAKES column but not in the MedLEE column.

2) Dissimilarity of number of unexplained symptoms

As mentioned above, cTAKES had 29 unexplained symptoms while MedLEE had only 23 such symptoms. We observed that the differences in annotation between two NLP engines is one of the reason for such discrepancy. The third row of Table IX shows such an example. The same concept is identified slightly differently by the two NLP engines. Our knowledge base contains the concept *soreness*, but not the concept *muscle pain*. Hence the concept *muscle pain* is not identified as unexplained in the MedLEE corpus. The differences in negation detection also cause the cTAKES corpus to have more unexplained symptoms than the MedLEE corpus.

3) Dissimilarity in the frequency of co-occurring disorders for unexplained symptoms (Refer Table II)

The dissimilarity in the frequency of co-occurring disorders in Table II for same disorder in the two corpora is due to the dissimilarity in the number of unexplained instances for *edema* in Table I. The cTAKES corpus has a significantly higher number of unexplained *edema* instances, hence it has a higher chance of co-occurring with any disorder.

4) Dissimilarity of the number of suggested relationships (Refer Table III)

The inequality of the number of suggestions as shown in Table III is due to two reasons: 1.) cTAKES has more unexplained symptoms and the suggestions generated by those 6 extra unexplained symptoms are present only in the cTAKES corpus, and 2.) the number of unexplained instances in the cTAKES corpus are higher than that in the MedLEE corpus. This means a symptom is found to be unexplained in a higher number of EMR documents. Hence this symptom co-occurs with a higher number of distinct disorders because each document may contain a different set of disorders. This causes it to have higher number of suggestions in the cTAKES corpus.

5) Dissimilarity in the increment of explanatory power (Refer Table VIII)

The difference in the increment of explanatory power is due to extra correct causal relationships found by the cTAKES corpus. Also, the higher number of unexplained instances in cTAKES helps to show a higher increment of explanatory power. For example, with the initial knowledge base, *edema* had 206 unexplained instances in the MedLEE corpus and 910 in the cTAKES corpus(Table I). Also, *hypertension* co-occurred 116 times and 647 times with those unexplained instances respectively (Table II). So when the relationship between *edema* and *hypertension* is discovered by our approach, the unexplained count is decreased by 206 in the MedLEE corpus and by 910 in the cTAKES corpus. Hence a higher increment of explanatory power is gained by the knowledge base relative to the cTAKES corpus than relative to the MedLEE corpus.

V. LIMITATIONS

As shown in the evaluation, the proposed algorithm has good precision in suggesting relationships. But it has the following limitations.

- It is unable to deal with complex relationships.
  Our knowledge base contains only single symptom to single disorder relationships, but it is possible that a single symptom can be explained by the existence of multiple disorders. This method is not able to capture such complex relationships.
- It may still miss potential relationships.
  The proposed algorithm might miss some causal relationships in EMR document due to two reasons: 1.) if the same symptom can be explained by multiple disorders in an EMR, we may not attribute the symptom to all of them, and 2.) if none of the neighbors of a co-occurring disorder has a relationship with the unexplained symptom, we may miss considering it as a candidate for suggesting a relationship.
- The precision of suggested relationships depends on the precision of the NLP engine.
  The proposed method requires the NLP engine to annotate the entities and associate negation and temporal information with the entity. The errors in the NLP output can affect the precision of proposed method.

VI. Conclusion

We have proposed a semantics driven semi-automatic method to improve the coverage and quality of an existing background knowledge base. The algorithm uses EMR data to identify the absence of causal relationships between symptoms and disorders in background knowledge and suggests plausible relationships that can rectify missing relationships using semantics of the domain concepts(existing causal relationships and hierarchical relationships). Our method minimizes the burden on domain experts by reducing the number of associations that they need to validate.

The proposed algorithm has better precision in suggesting plausible relationships compared to a simpler co-occurrence based method, and holds the promise for good recall given more data. The co-occurrence based method may have a better recall at the cost of precision, since it does not suffer from the deficiency caused by the neighbors collecting step of the proposed algorithm. In summary, the algorithm enables making effective use of domain experts for building high quality knowledge bases.

VII. Acknowledgement

We would like to thank the domain experts Anjali Rami and Deval Pathak who helped us in validating the suggested relationships by proposed method.

References

[1] K. Anyanwu, A. Sheth. "The P Operator: Discovering and Ranking Associations on the Semantic Web." SIGMOD Record (Special issue on Amicalola Workshop), 31 (4), pp. 42-47, December 2002.

[2] C. Freeman, P. Alderson, J. Austin, J. Cimino, S. Johnson. "A general natural-language text processor for clinical radiology." Journal of the American Medical Informatics Association, vol. 1, pp 161-174, 1994.

[3] A. Aronson. "Metamap: Mapping text to the umls metathesaurus." Bethesda, MD: NLM, NIH, DHHS, 2006.

[4] B. Aker. "Emerging Semantic Web Technology Could Help Intelligence Analysts Spot New Terror Threats." http://www.govtech.com/pcio/Semantic-Web-Could-Help-Spot-Terror-Threats-021111.html

[5] J. Zaino. "Semantic Web Tech: Can It Make "Top Secret America" A More Secure America." http://semanticweb.com/semantic-web-tech-can-it-make-top-secret-america-a-more-secure-america_b709

[6] A. Sheth, I. Arpinar, V. Kashyap. "Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships." Enhancing the Power of the Internet (Studies in Fuzziness and Soft Computing), vol. 139, pp. 63-94. 2004.

[7] S. Schulz, R. Cornet. "SNOMED CT's Ontological Commitment." In Proc. ICBO: International Conference on Biomedical Ontology; National Center for Ontological Research, 2009, pp. 55-58.

[8] O. Bodenreider. "The Unified Medical Language System (UMLS): integrating biomedical terminology." Nucleic Acids Res 2004;32:D267-D270.

[9] S. Perera, C. Henson, K. Thirunarayan, A. Sheth. "Data Driven Knowledge Acquisition Method for Domain Knowledge Enrichment in the Healthcare." 6th International Conference on Bioinformatics and Biomedicine BIBM12, Philadelphia, 4-7 Oct, 2012, pp. 197

[10] C. Henson, K. Thirunarayan, A. Sheth. "An Ontological Approach to Focusing Attention and Enhancing Machine Perception on the Web." Applied Ontology, vol. 6(4), pp. 345-376, 2011.

[11] P. Desai, C. Henson, P. Anatharam, A. Sheth. "Demonstration: SECURE – Semantics Empowered resCUe Environment." In Proc 4th Intl. Workshop on Semantic Sensor Networks, pp. 110-113, co-located with the 10th International Semantic Web Conference (ISWC 2011), 23-27 October 2011, Bonn, Germany, 2011.

[12] G. Savova, J. Masanz, P. Ogren, et al. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." Journal of the American Medical Informatics Association. 2010 Sep 1;17(5):507-13. 2010

[13] M. Sabou, M. D'Aquin, E. Motta. "Exploring the Semantic Web as Background Knowledge for Ontology Matching." Journal on Data Semantics XI, vol. 5383, pp. 156-190, 2008

[14] F. Zablith. "Evolva: A comprehensive approach to ontology evolution." In Proc 6th European Semantic Web Conference (ESWC) PhD Symposium, 2009, pp 944-948.

[15] D. Faure, C. Nedellec. "A corpus-based conceptual clustering method for verb frames and ontology acquisition." In Proc LREC workshop on adapting lexical and corpus resources to sublanguages and applications, pp. 5-12, 1998.

[16] M. Kavalec, A. Maedche, V. Svatek. "Discovery of Lexical Entries for Non-taxonomic Relations in Ontology Learning." SOFSEM Theory and Practice of Computer Science, vol. 2932, pp. 17-33, 2004

[17] M. Ciaramita, A. Gangemi, E. Ratsch, S. Jasmin, R. Isabel. "Unsupervised Learning of Semantic Relations between Concepts of Molecular Biology Ontology." In proc International Joint Conference on Artificial Intelligence, pp. 659-664, 2005.

[18] A. Schutz, P. Buitelaar. "RelExt: A Tool for Relation Extraction from Text in Ontology Extension." In Proc 4th International Semantic Web Conference, pp. 593-606, 2005.

[19] D. Sanchez, A. Moreno. "Discovering Non-taxonomic Relations from the Web." In: Corchado, E.S., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 629-636. Springer, Heidelberg (2006)

[20] C. Xiao, D. Zheng, Y. Yang, G. Shao. "Automatic Domain-Ontology Relation Extraction from Semi structured Texts," Asian Language Processing, 2009. IALP '09. International Conference on , vol., no., pp.211-216, 7-9 Dec. 2009

[21] Q. X. Do, Y. S. Chan, D. Roth. "Minimally supervised event causality identification." In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), pages 294-303.

[22] E. Blanco, N. Castell, D. Moldovan. "Causal relation extraction." In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), May 2008.

[23] C. Khoo, S. Chan, Y. Niu, A. Ang. "A method for extracting causal knowledge from textual databases." Singapore Journal of Library & Information Management, 28, 48-63.

[24] R. Girju, D. I. Moldovan, "Text Mining for Causal Relations." Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference, p.360-364, May 14-16, 2002

[25] R. Mulkar-Mehta, C. Welty, J. R. Hobbs, E. Hovy. "Using Granularity Concepts for Discovering Causal Relations." (2011).

[26] C. Friedman, L. Shagina, Y. Lussier, G. Hripcsak. "Automated encoding of clinical documents based on natural language processing." Journal of American Medical Informatics Association, pp 392-402, 2004