Implicit Concept Recognition in Healthcare Data Augmented by External Resources

Subiksha.K. P
Research Scholar
School of Information Technology
Madurai Kamarajar University
Palkalai Nagar,Madurai, India

M. Ramakrishnan M.E,Ph.D,Ph.D
Professor and Head, Chair Person
School of Information Technology
Madurai Kamarajar University
Palkalai Nagar,Madurai, India

Abstract **- Enquiries and documents were reforged from their term-based originals into medical concepts as outlined by the Symbolic Nomenclature Of Medicine – Clinical Terms ontology. Analysis on a real-world assortment of medical records showed that clinical-terminology or concept with contextual ontology used approach surpassed the keyword baseline in Mean Average precision. Additionally, the concept with contextual ontology based approach created important enhancements on exhausting queries. The proposed concept and contextual-based on user role approach provides a system for additional development into reasoning primarily based search systems for managing healthcare knowledge. An approach of searching EMR that is based on concepts and contextual search based on user role on the contrary of keyword matching is tried.**

Index Terms - Concept-based Information Retrieval, ATOMS (AGENT based Terminology Management System with Ontology Representation), ICR, WSD, CRT.

I. INTRODUCTION

Electronic medical record (EMR) data are becoming increasingly important for quality improvement, comparative effectiveness research, evidence-based medicine, and establishing robust phenotypes for genomic analysis. Unfortunately, most EMR implementations were designed to facilitate one-on-one interactions, not to support analysis of aggregated data as required by many secondary uses. As a result, efforts to repurpose clinical data must contend with few widely implemented data standards and large amounts of potentially useful information stored as unstructured free text. Concept-based Information Retrieval (CBIR) intended to make use of happening knowledge sources in order to render further information and set of facts that may not be declared in a document aggregation and users queries.
Ontologies can be used for case representation and storage, and it can be used as background knowledge. Using standard medical ontologies, such as SNOMED CT, enhances the interoperability and integration with the health care systems.

Data mining (DM) and machine learning (ML) methods provide an opportunity for researchers to discover Implicit Concept Recognition (ICR) regularities buried in the large volume of clinical records. There has been some work on this problem. Existing methods have been validated on a limited amount of manually well-structured data. However, the contents of most fields in the clinical records are unstructured. As a result, the previous methods verified on the well-structured data will not work effectively on the Electronic Medical records, and the EMRs are, consequently, required to be structured in advance. Manually structuring the large volume of EMRs is time-consuming and labour-intensive, but the development of automatic methods for the structuring task is at an early stage. Therefore, in this paper, Implicit Concept recognition (ICR) in the chief complaints, which is one of the important tasks to structure the ICRs of clinical text, is carefully studied.

Keyword searches often return a result that includes large number of false positives or that exclude too many false negatives because of the effects of synonymy and polysemy[7]. Synonymy means that one of two or more words in the same language have the same meaning, and polysemy means that many individual words have more than one meaning [8]. In addition to the problems of polysemous and synonymy, keyword searches can exclude inadvertently misspelled words as well as the variations on the stems of words. Keyword searches are also susceptible to errors introduced by optical character recognition (OCR) scanning processes, which can introduce random errors into the text of documents during the scanning process.

II. BACKGROUND

The application motivating this study is the retrieval of relevant documents from EMR systems. The identification of relevant documents is a prerequisite to most secondary data uses, such as automated quality measurement, medical record-based research, cohort identification, and comparative effectiveness research. Unfortunately, queries of structured data fields such as ICD-9 codes and Current Procedural Terminology (CPT) codes for secondary data use have proven less than ideal. The questionable quality of administrative code assignments has been documented extensively since the rise of administrative code-based reimbursement, and custom case-finding algorithms can be time consuming to develop and must be evaluated for each application. A solution to this dilemma may be provided by clinical IR technologies.

In the past two decades, clinical IR has evolved from a field with few researchers working on even fewer systems to the release of open-source components and libraries. More recently, researchers in the fields of computer science and linguistics have released open-source software frameworks upon which IR methods can be developed.

Clinical IR researchers have capitalized on these frameworks, producing modular pipelines for specific retrieval applications.

One such pipeline for clinical NLP is the Clinical Text Analysis and Knowledge Extraction System (cTAKES). The cTAKES maps free text to SNOMED concepts and is based on the open-source Unstructured Information Management Architecture (UIMA). Narrative text is the primary communication method in the medical domain. Much of the important patient information is only found in clinical notes in electronic medical records. NLP technologies offer a solution to convert free text data into structured representations. Over the last two decades, there have been many efforts to apply NLP technologies to clinical text. The Linguistic String Project, the Medical Language Extraction and Encoding System (MedLEE) and SymText/MPlus are a few of the earliest NLP systems developed for clinical domain. More recently, open source clinical NLP systems such as cTAKES and HiTEX have also been introduced into the community. Most of clinical NLP systems can extract various types of named entities from clinical text and link them to concepts in the Unified Medical Language System (UMLS), such as MetaMap and KnowledgeMap. After a clinical concept is identified, many applications require determination of its assertion (ie, is a medical condition present or absent?). Among various negation detection algorithms, NegEx[13] has arguably been used the most widely and has been incorporated into many systems.

III. FEATURE SELECTION

Instead of using every possible feature for our classifier, or manually selecting our set of features through trial and error, we use an automated feature selection approach to finding the best set of features. Because our feature set is chosen automatically, we refer to our approach as having a flexible ATOMS architecture. Given a new task, or simply new data, we can automatically determine a new set of features so long as the new task operates on the same type of input. For example, classifying a concept's type (problem, test, treatment) and a concept's assertion type (present, absent, etc) both operate on the concept level. In both of these tasks, we made largely the same set of features available to the feature selector.

EXTERNAL RESOURCES

We use numerous external resources to derive features. These resources include UMLS, MetaMap, NegEx, GENIA, WordNet, PropBank, the General Inquirer, and Wikipedia. various types of features that were extracted from the word itself and its context, including:

- A.Word Level Information: Bag-of-word, Orthographic information—such as capitalization of letters in words, and prefixes and suffixes of words;
- B.Syntactic Information: Part of Speech tags obtained using MedPOST (http://www.ncbi.nlm.nih.gov/staff/lsmith/MedPost.html);
- C.Lexical and Semantic Information from NLP systems: mainly normalized concepts (eg, UMLS concept unique identifiers) and semantic types identified by NLP systems. Three NLP systems were used: (1) MedLEE; (2) KnowledgeMap; (3) a Dictionary-based Semantic Tagger (DST) developed for this task, which uses vocabularies from public (eg, UMLS) and private (eg, Vanderbilt's problem list) sources and filtered them for medical problems, tests, and treatments;
- d.Discourse Information: Sections in the clinical notes (eg, 'Current Medications' section) and Sources of the notes (eg, 'Partners HealthCare System'), obtained by customized programs developed for the challenge data.

PARSER

To acquire the empirical evidence of the usefulness of sentential syntax, we parsed the input text using the Charniak's ME reranking parser [3] with its improved, self-trained biomedical parsing model [11]. These were then converted into Stanford dependencies [6].4 The features we extracted from the dependency parsing trees included words, their dependency tags, and arc labels on the dependency path between the two minimal trees that cover each of the two concepts, respectively, along with the word type and tags of their common ancestor, as well as the minimal, average and maximal tree distances between these two minimum-covering trees and their common ancestor.

UMLS/METAMAP. The first explicit, manually created knowledge base we incorporate is the Unified Medical Language System (UMLS) [17], created and maintained by the U.S. National Library of Medicine (NLM) to ''facilitate the development of computer systems that behave as if they 'understand' the meaning of the language of biomedicine and health.'' This knowledge base contains a unified thesaurus and ontology, the mapping between different terminology systems and disparate databases, as well as the corresponding software tools that perform on these data. The UMLS Meta-thesaurus covers over 1 million biomedical concepts and 5 million concept names, and was created from more than 100 different

vocabulary sources with human intervention of editing and reviewing. Specifically in this study, we applied MetaMap [1], which is a widely-used entity recognition tool in the biomedical domain. We used MetaMap to recognize lexical variations of medical concepts from UMLS within their context. With the MetaMap matching results, we can represent words by their domain-specific semantic categories, i.e., UMLS semantic types such as ''sign or symptom'' and ''therapeutic or preventive procedure''. These labels are used as features to hopefully smooth the sparseness of lexicalized features. The semantic-type labels are associated with words in this system. when MetaMap assigns a label to a multi-word phrase, we break the phrase into words and assign the same label to each word to acquire flexibility in feature construction. More specifically, we use the unigram UMLS labels of the three words before and after the two concepts in

question, of the words between them and of the words contained in them. In addition, we use UMLS label pairs associated with each word pair from the two concepts, i.e., one label from each concept.

Domain word/phrase clusters. We have also manually created word/phrase clusters specifically for clinical text to further smooth data sparseness. For example, we created a list to include words, phrases, and doctors' shorthands that express indication such as ''p/w'', ''have to do with'', '' secondary to'', '' assoc w/''. Another example is a resistance list containing words/phrases such as ''unresponsive'', ''turn down'', and ''hold off''. We extracted these features in a way similar to that described in Section 4.1. That is, we identify if words in a domain-word list appear within three words before or after the two concepts in question, and extract these as binary features. Similarly, we check the occurrences of these domain word/phrase lists among words between the two concepts and words in the concepts. An example of such features is: ''a resistance word appears within three words after a problem.''

Automatically-acquired domain knowledge In addition to the explicit domain semantics that are created manually, there is abundant domain knowledge embedded in much larger free-text. We are also curious about its usefulness in this task. MEDLINE, for example, is a bibliographic database of life sciences and biomedical information. It includes 5000 selected resources and covers such publications from 1950s to the present, including health-related fields such as medicine, nursing, pharmacy, dentistry, veterinary medicine, preclinical sciences, and healthcare. The database contains more than 18 million records approximately and has been widely used in various healthcare related research. In this work, we calculate the pointwise mutual information (PMI) between the two given concepts in all the abstracts of MEDLINE articles to estimate their relatedness. The motivation is that such information could provide evidence to help determine the likelihood that the two concepts have a positive relation, though not necessarily their specific relation categories.

IV. CONCEPT EXTRACTION

The overall architecture of our concept extraction approach is shown in figure 1. Each discharge summary in the dataset is provided with tokenization and sentence boundaries. We use regular expressions to recognize nine entity types that support concept extraction: names, ages, dates, times, IDC identifiers, percents, measurements, dosages, and list elements. Each sentence is then categorized as being prose or non-prose using a simple heuristic. Sentences that end with a colon are assumed to be section headers and are not considered prose. A sentence is considered prose if it ends with a period or question mark, or if it consists of at least five tokens, less than half of which may be punctuation. Otherwise, it is considered non-prose. We then detect concept boundaries (start and end tokens) using two CRF classifiers: one CRF for prose sentences; the other CRF for non-prose sentences. For concept extraction, we used only greedy forward feature selection. The feature selector primarily chose lexical and pattern-entity features for non-prose concepts, along with MetaMap features. For prose concepts, a wide variety of features commonly used in NLP were chosen, including the four annotations provided by GENIA. The lists of features chosen by the feature selector for each CRF classifier are shown in Table 1.

After detecting the concept boundaries, our approach classifies each concept as a problem, treatment, or test. We use a single SVM classifier for all concepts, prose and non-prose, and employ the same greedy feature-selection technique. The selected features are shown in Table 1.

The feature selector for boundaries chose from a set of 125 features, choosing seven for non-prose concept boundaries and 15 for prose concept boundaries. For concept type, a total of 222 features were available to the feature selector (most of which were developed for assertion classification), of which eight were chosen. For features that can take non-numeric values (eg, NF1 can be any word, while TF1 can be many words for a given concept), we expand these features into N binary features, where N is the number of values seen for the feature in the training data. This results in large, sparse feature vectors that can be problematic for some machine-learning techniques, but are easily handled by SVMs and CRFs. Clinical records contain information that can be invaluable, for example, for pharmacovigilance, for comparative effectiveness studies, and for detecting adverse events. The structured and narrative components of clinical records collectively provide a comprehensive account of the medications of patients. The medication challenge was designed as an information extraction task. The goal, for each discharge summary, was to extract the following information (called 'fields'5) on medications experienced by the patient:

- Medications (m): including names, brand names, generics, and collective names of prescription substances, over the counter medications, and other biological substances for which the patient is the experiencer.
- Dosages (do): indicating the amount of a medication used in each administration.
- Modes (mo): indicating the route for administering the medication.
- Frequencies (f): indicating how often each dose of the medication should be taken.
- Durations (du): indicating how long the medication is to be administered.
- Reasons (r): stating the medical reason for which the medication is given.
- List/narrative (ln): indicating whether the medication information appears in a list structure or in narrative running text in the discharge summary.

Line no.	text
63	well. Although left transmetatarsal amputation being considered ,
64	it was felt that she had a good chance of healing the wound
65	appropriately. She had a single temperature spike , although all
66	cultures remained negative. *She had continuation of her Heparin*
67	*while she was started on a course of Coumadin to reserve patency of*
68	*her graft. …*

Gold standard

m="heparin" 66:8 66:8||do="nm" ||mo="nm" ||f="nm" ||du="nm" ||r="nm"||ln="narrative"
m="coumadin" 67:8 67:8||do="nm" ||mo="nm" ||f="nm" ||du="nm" ||r="her graft." 68:0 68:1||ln="narrative"

Table 1 - Sample Narrative Text

V. AGENT BASED TERMINOLOGY MANAGEMENT SYSTEM WITH ONTOLOGY REPRESENTATION

Healthcare information is available in various disparate systems, so the Agent based system is implemented and Ontologies that promotes shared understanding of Terminologies and it determines a novel ontology for representing the medical domain, based on concepts search in standard medical ontologies [7].

The system has the following four main phases:-

1. Concept Identifier and Mapping Phase
2. Contextual Phase
3. OntoMap Phase
4. Evaluation and Retrieval Phase

An Ontology based Query expansion is done to improve the precision-recall of the search results by concentrating on the context of concept(s). The relevant k-cores are matched with the ontology of medical domain to extract the concepts based on the similarity measure. The most relevant concepts along with the ranked k-cores are selected based on the preferences of the user which was mentioned in user profiles. The user query is enriched with the selected concept and passed to the search engine for efficient retrieval of relevant documents. Relevance feedback is used in case the query need to be refined or else the intelligent Word Sense Disambiguation (WSD) would retrieve the relevant results with high precision and recall values.

The documents are processed and concepts are extracted. Find relationships among concepts and ontology is constructed. For the existing EMR knowledge bases missing relationships are found. New relationships are identified. Validating the existing dataset as well identified. ATOMS resolves terminologies ambiguities, polysemy and synonymy problems that exist in keyboard baseline retrieval models. Ontologies promote shared understanding of Terminologies by various users in different roles. In this system Data driven paradigm have been proposed. Data driven method means program statements describe the data to be matched and the processing required rather than defining a sequence of steps to be taken. The data driven method assumed that each symptom in an Electronic Medical Record (EMR) document should be explained by at least one disorder present in the document

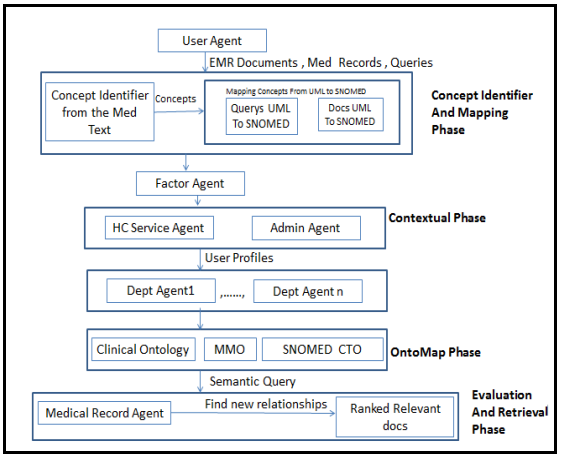

Figure 1 - ATOMS Architecture.

.

At the top of the architecture is placed the user, who interacts with the system through his User Agent (UA). This agent stores static data related to the user and dynamic data. The Factor Agent (FA) is an agent that knows about all the medical services as well the Admin agent assigns role for the user. That includes Contextual phase. Each department has a staff of several doctors, modelled through Department Agents (DAs), and offers more specific services, also modelled as SAs. At the bottom of the architecture, a Medical Record Agent (MRA) controls the access to a database that stores all EMR of the patients of the medical centre [12]. Appropriate security measures have been taken to ensure that only properly authenticated and authorised agents may access and update the EMR.

One relevant study investigated the contribution of syntactic information to semantic categorization of words in discharge summaries using Support Vector Machines (SVM); but the study was done on a small data set with 48 clinical notes. In this paper, we describe a systematic investigation on ML-based approaches for recognizing broad types of clinical entities and determining their assertion status, and report a new hybrid clinical entity extraction framework, which achieved good performance.

VI. IMPLICIT CONCEPT RECOGNITION

Terminology refers to a system of words used to name things in a particular discipline. Terminologies define the meaning of data (meaning) i.e. changes data to information through instantiation of semantic rules.

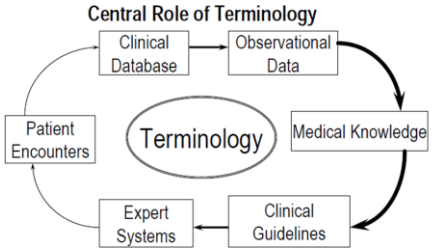

Figure 2 – Role Of Terminology

With the increasing automation of health care information processing, extraction of meaningful information from textual notes in electronic medical records (EMR) has become critical. One of the key challenges is extraction and normalization of concepts mentions. State-of-the-art approaches have focused on the recognition of concepts explicitly mentioned in EMR. However, clinical documents often contain phrases that indicate concepts but do not contain their names. Considered those implicit concepts mentions and introduce the problem of implicit Concept recognition (ICR) in clinical documents. The solution has been proposed to ICR that leverages concepts definitions from a knowledgebase to create concepts models, projects sentences to the concepts models and identifies implicit concepts mentions by evaluating semantic similarity between sentences in clinical documents and concepts models.

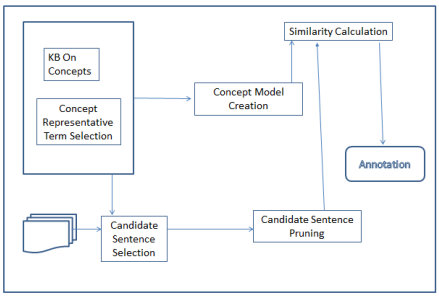

Figure 3 - Components of the Proposed Solution

The above figure shows the components of the solution which are discussed below in detail. In order to facilitate the sub-tasks, the algorithm introduces the concept of concept representative term for each concept and proposes an automatic way to select these terms from concept definitions.

A. Concept representative term (CRT) selection finds a term with a high representative power to concept and plays an important role in defining it[8]. The representative power of a term t for concept c is defined based on two properties: its dominance among the definitions of concept c, and its ability to discriminate the mentions of concept c from other concepts. This is formalized in eq. (1). Consider the concept 'appendicitis' as an example. It is defined as 'acute inflammation of appendix'. Intuitively, both terms inflammation and appendix are candidates to explain the concept appendicitis. However, the term appendix has more potential to discriminate the implicit mentions of appendicitis than the term inflammation, because the term inflammation is used to describe many concepts. Also, none of the definitions define appendicitis without using the term appendix; therefore, appendix is the dominant term, and consequently it has the most representative power for the concept 'appendicitis'. Using a score inspired by the TF-IDF measure to capture this intuition. The IDF (inverse document frequency) value measures the specificity of a term in the definitions. The TF (term frequency) captures the dominance of a term. Hence the representative power of a term t for concept c (*rt*) is defined as,

$$rt = freq(t,Qc) * log \frac{|C|}{|Ct|}$$

Qc is the set of definitions of concept c, C is the set of all concepts. freq(t,Qe) is the frequency of term t in set Qc, |C| is the size of the set C (3962 in our corpus), and the denominator |Ct| calculates the number of concepts defined using term t. Expanding the CRT found for the concept with this technique by adding its synonyms obtained from WordNet.

B. Concept Model Creation

Our algorithm creates concept indicator from a definition of the concept. A concept indicator consists of terms that describe the concept. Consider the definition 'A disorder characterized by an uncomfortable sensation of difficulty breathing' for 'shortness of breath', for which the selected CRT is 'breathing'. The terms uncomfortable, sensation, difficulty, and breathing collectively describe the concept. A negative addition of other terms to this definition of the concept indicator affects the similarity calculation with the candidate sentences since they are less likely to appear in a candidate sentence.

C. Candidate Sentence Selection

The sentences with CRT in an input text are identified as candidate sentences containing implicit mention of the corresponding concept. A sentence may contain multiple CRTs and consequently become a candidate sentence for multiple concepts. This step reduces the complexity of the classification task as now a sentence has only a few target concepts.

D. Candidate Sentence Pruning

In order to evaluate the similarity between any given candidate sentence and the concept model, perform a projection of candidate sentences onto the same semantic space. Can implement this by pruning the terms in candidate sentences that does not participate in forming the segment with implicit concept mentions. Candidate sentences are pruned by following the same steps followed to create the concept indicators from the concept definitions.

E.Semantic Similarity Calculation

As the last step, the proposed solution determines the similarity between the concept model and the pruned candidate sentence. The sentences with implicit concept mentions often use adjectives and adverbs to describe the concept and they may indicate the absence of the concepts using antonyms or explicit negations. These two characteristics pose challenges to the applicability of existing text similarity algorithms such as MEDICAL CLASSIFICATION SYSTEM (Mihalcea et al., 2006) and matrixJcn (Fernando and Stevenson, 2008) which are proven to perform well among the unsupervised algorithms in paraphrase identification task (ACLWiki, 2014). Unfortunately, adjectives and adverbs are not arranged in a hierarchy, and terms with different part of speech (POS) tags cannot be mapped to the same hierarchy. Hence, they are limited in calculating the similarity between terms of these

categories. This limitation negatively affects the performance of ICR as the concept models and pruned sentences often contain terms from these categories. Consider the following examples:

1. Her breathing is still uncomfortable adjective.
2. She is breathing comfortably adverb in room air.
3. His tip of the appendix was inflamed verb.

The first two examples use an adjective and an adverb to mention the concept 'shortness of breath' implicitly. The third example uses a verb to mention the concept 'appendicitis' implicitly instead of the noun inflammation that is used by its definition, developing a text similarity measure to overcome these challenges and weigh the contributions of the words in the concept model to the similarity value based on their representative power.

F. Handling Negations

Negations are of two types:

1) Negations mentioned with explicit terms such as no, not, and deny, and

2) Negations indicated with antonyms (e.g., 2nd example in above list).

NegEx algorithm (Chapman et al., 2001) is used to address the first type of negations. Addressing the second type of negations, needs exploitation of the antonym relationships in the WordNet. The similarity between the concept model and the pruned candidate sentence is determined by computing the similarities of their terms. The term similarity is computed by forming an ensemble using the standard WordNet similarity measures namely, WUP, Resnik (Resnik, 1995), LIN (Lin,1998), JCN (Jiang and Conrath, 1997), as well as a predict vector-based measure Word2vec (Mikolov et al., 2013) and a morphology-based similarity metric. Levenshtein1 as:

sim(t1, t2) = max m€M (simm(t1, t2))

where t1 and t2 are input terms and M is the set of the above mentioned similarity measures. This ensemble-based similarity measure exploits orthogonal ways of comparing terms: semantic, statistical, and syntactic. An ensemble-based approach is preferable over picking one of them exclusively since they are complementary in nature, that is, each outperforms the other two in certain scenarios. The similarity values calculated by WordNet similarity measures in *simm(t1, t2)* are normalized to range between 0 and 1. The similarity of a pruned candidate sentence to the concept model is calculated by determining its similarity to each concept indicator in the concept model, and picking the maximum value as the final similarity value for the candidate sentence. The similarity between concept indicator e and pruned sentence s, *simm(c, s)* is calculated by summing the similarities calculated for each term tc in the concept indicator weighted by its representative power as defined in *rt*. If tc is an antonym for any term in s (ts), it contributes negatively to the overall similarity value, else it contributes to the linear portion of the maximum similarity value between tc and some ts . The overall similarity value is normalized based on the total representative power of all the terms tes and ranges between -1 and +1.

$$Sim(c, s) = \frac{\sum_{tc \in C} f(tc,s) * r_{tc}}{\sum_{tc \in C}}$$

Note that this formulation weighs the contribution of each term according to its importance in defining the concept. The higher similarity with a term that has higher representative power leads to higher overall similarity value, while the lower similarity with such terms leads to a lower total similarity value.

$$f(tc, s) = \begin{cases} -1 & \propto x(tc, s) == 0 \\ max_{ts \in S} \mathrm{Sim}(tc, ts) & otherwise \end{cases}$$

The task of CT standardization is a combination of WSD and semantic similarity where a term is mapped to a unique concept in an ontology which is based on the description of that concept in the ontology after disambiguating potential ambiguous surface words, or phrases [10-11]. This is especially consistent for abbreviations and acronyms which are much more common in healthcare information (Moon et al., 2012).

VII.EVALUATION

Two supervised machine-learning classifiers used in this study are maximum entropy (MaxEnt) and conditional random fields (CRFs). MaxEnt is a framework for estimating probability distributions from a set of training data. Maximum entropy models have been used in NLP to chunk phrases,24 for part-of-speech tagging, and in a number of biomedical applications.

A CRF is an undirected graphical model with edges representing dependencies between variables. Peng and McCallum showed that CRFs outperform the more commonly used support vector machines in extracting common fields from the headers and citations of literature. Wellner etal showed the ability of CRFs to achieve high levels of performance in the deidentification of personal health identifiers, limiting customization to manual annotation of training sets.

A supervised machine learning approach to discover relations between medical problems, treatments, and tests mentioned in electronic medical records. A single support vector machine classifier was used to identify relations between concepts and to assign their semantic type. Several resources such as Wikipedia, WordNet, General Inquirer, and a relation similarity metric inform the classifier. Variation on the Surgical Treatment of Early Stage Breast Cancer. This study investigates institutional variance in the surgical processes of breast cancer surgical care. Collaboration with Dana-Farber Cancer Institute, Institute for Health Metrics and the University of Wisconsin.

VIII. RESULTS

All free text medical records for patients in 68 community hospitals are used. Sample size is of about 1300 randomly sampled documents from all records. It identifies Ischemic Stroke in the VHA. It consists of Text notes, discharge summaries and consults from the VA. Four types of

reports where found in the corpus: 61 discharge summaries, 54 ECG reports, 42ECHO reports and 42 radiology reports, for a total of 199 training documents, each containing several disorder mentions. The annotation focus was on disorder mentions, their various attributes and normalizations to an UMLS CUI.

Table 2 - Sample Data set

Dataset Types	Type	Note	Concept	Concept Id	CUIless
Training Data	ALL	199	5816	4177	1639
	Echocardiogram	42	828	662	166
	Radiology Rep	42	555	392	163
	Discharge summaries	61	3589	2646	943
	Electrocardiogram	54	193	103	90
Dev-Data	ALL	99	5340	3619	1721
	Echocardiogram	12	338	241	97
	Radiology Rep	12	162	126	36
	Discharge summaries	75	4840	3252	1588
	Electrocardiogram	0	0	0	
Test-Data	ALL		133	-	-

A concept was in the Disorder semantic group if it belonged to one of the following UMLS semantic types: Congenital Abnormality; Acquired Abnormality; Injury or Poisoning; Pathologic Function; Disease or Syndrome; Mental or Behavioural Dysfunction; Cell or Molecular Dysfunction; Experimental Model of Disease; Anatomical Abnormality; Neoplastic Process; and Signs and Symptoms.

Table 3 - MAP and Precision Measures

Approaches	Mean Avg Precision	Precision	Recall
Breast cancer Operative Reports – Keyword Baseline Approach	0.95	0.98	0.98
Breast cancer Operative Reports Concept-based Approach	0.99	0.97	0.89
Breast cancer Operative Reports Role and Ontology based Approach	0.85	0.91	0.87
Breast cancer Clinical Notes – Keyword Baseline Approach	0.97	0.96	0.98
Breast cancer Clinical Notes Concept-based Approach	0.93	0.97	0.89
Breast cancer Clinical Notes Role and Ontology based Approach	0.89	0.86	0.95
Breast cancer Pathology Reports – Keyword Baseline Approach	0.93	0.97	0.99
Breast cancer Pathology Reports Concept-based Approach	0.96	0.99	0.95
Breast cancer Pathology Reports Role and Ontology based Approach	0.91	0.92	0.92

A disorder mention was defined as any span of text which can be mapped to a concept in SNOMEDCT and which belongs to the Disorder semantic group. It also provided a semantic network in which every concept is represented by its CUI and is semantically typed (Bodenreider and Mc- Cray, 2003). The Finding semantic type was left out as it is very noisy and our pilot study showed lower annotation agreement on it. Following are the salient aspects of the guidelines used to annotate the data. Annotations represent the most specific disorder span. For example, small bowel obstruction is preferred over bowel obstruction.

On top of that, a formal evaluation of the contextualization techniques may require a significant amount of extra feedback from users in order to measure how much better a retrieval system can perform with the proposed techniques than without them.

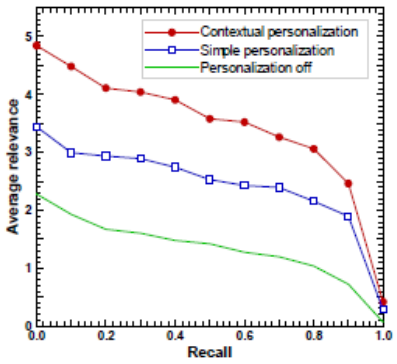

Figure 4 - Comparative performance of personalized search with and without contextualization averaged over ten use cases

It is necessary to compare the performance of retrieval a) without personalization, b) with simple personalization, and c) with contextual personalization. In this case, the standard evaluation measures from the IR field require the availability of manual content ratings with respect to a) query relevance, b) query relevance and general user preference (i.e. regardless of the task at hand), and c) query relevance and specific user preference (i.e. constrained to the context of his/her task). Ontologies can represent crucial information when building WSD systems, for two main reasons: i) ontologies distinctively organizes the most important terms of a scientific domain and they would help to build more exerting context vectors based on ontological concepts in the final outcome and ii) the structure of the ontology can be strategically used to devise new techniques for Word sense Disambiguition..

REFERENCES

[1] W.R. Braithwaite, "The federal role in setting standards for the exchange of health information", Proceedings of the Symposium on Pacific Medical Technology (PACMEDTEK '98). IEEE Computer Society, Washington, D.D., USA, 2012, pp. 340-347.

[2] M. Vida, O. Lupse, L. Stoicu-Tivadar, "Improving the interoperability of healthcare information systems through HL& CDA and CCD

standards", 7th IEEE International Symposium on Applied Computational Intelligence and Informatics. Timisoara, Romania 2012. pp. 157-161.

[3] Voorhees EM. In: Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval. Dublin, Ireland: ACM; 1994.

[4] Fellbaum C. WordNet: An electronic lexical database. Cambridge, MA.: The MIT Press; 1998.

[5] Ravindran D, Gauch S. In: Proceedings of the 13th annual international ACM CIKM conference on in-formation and knowledge management. ACM; 2004. Exploiting hierarchical relationships in conceptual search. pp. 238–239.

[6] Liu Z, Chu WW. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. Information Retrieval. 2007 Jan;10(2):173–202.

[7] Steindel S.J. Liu Z, Chu WW. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. Information Retrieval. 2007 Jan; 10(2):173–202.

[8] Zheng HT, Borchert C, Jiang Y. A knowledge-driven approach to biomedical document conceptualization. Artificial Intelligence in Medicine. 2010; 49(2):67–78."

[9] Z. Li, et al, "A secure electronic medical record sharing mechanism in the cloud computing platform", IEEE 15th International Symposium on Consumer Electronics, Singapore, 2011, pp. 98-103.

[10] A. T. Swartout. Ontologies. IEEE Intelligent Systems and Their Applications, 14(1):18–19, 1999.

[11] R. B. Altman, M. Bada, and X. J. Chai. Riboweb: an Ontology-based system for collaborative molecular biology. IEEE Intelligent Systems and Their Applications, 14(5):68–76, 1999.

[12] N. Goga, S. Costache, F. Moldoveanu, "A formal analysis of ISO/IEEE P11073-20601 standard of medical device communication", 3rd Annual IEEE International Systems Conference, Vancouver, Canada, 2009. pp. 163-166.

[13] Hersh [Moreno et al., 2003a] Moreno, A., Isern, D., and Sanchez, D. (2003a). Provision of agent-based health care services. AI Communications. Special Issueon Agents in Healthcare, 16:135

[14] W. 3rd. New York: Springer Verlag: 2009. Information retrieval: a health and biomedical perspective.

[15] L. Yang, Y. Gu, "Design and realization of DICOM/HL7 gateway in PACS", IEEE 2011 International Conference on Electronic & Medical Engineering and Information Technology, Shanghai, China, 2011. pp. 2164-2167.

[16] T. Namli, G. Aluc, A. Dogac, "An interoperability test framework for HL7-based systems", IEEE Transactions on Information Technology In Biomedicine, vol. 13, no. 3, May 2009. pp .389-399.

[17] Reynolds, R.G., and Rychtyckyj, N.*, "Using Cultural Algorithms to Improve Performance in Semantic Networks", in Proceedings 1999 IEEE Congress on Evolutionary Computation, Washington, D. C., July 6-9, 1999, pp. 1651-1656.

[18] Reynolds, R.G., and Ostrowski, D.*, "Knowledge-Based Software Testing Agent Using Evolutionary Learning with Cultural Algorithms", in Proceedings 1999 IEEE Congress on Evolutionary Computation, Washington, D. C., July 6-9, 1999, pp. 1657-1663.

[19] Department of Health and Human Services. "HIPAA Administrative Simplification: Modifications to Medical Data Code Set Standards to Adopt ICD-10-CM and ICD-10-PCS."

[20] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010; 17(3):229–236.

[21] 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

[22] Nadkarni P, Marenco L. Implementing description-logic rules for SNOMED-CT attributes through a table-driven approach. J Am Med Inform Assoc 2010;17:182–4.

[23] Medical Subject Headings. National Library of Medicine. http://www.nlm.nih.gov/mesh/

[24] 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.

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

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

[27] 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