Bulletin of Environment, Pharmacology and Life Sciences
Bull. Env. Pharmacol. Life Sci., Vol 5 [5] April 2016: 01-07

Online ISSN 2277-1808
Journal's URL:http://www.bepls.com
CODEN: BEPLAD
Global Impact Factor 0.533
Universal Impact Factor 0.9804

ORIGINAL ARTICLE **OPEN ACCESS**

A Comprehensive Approach for Knowledge Acquisition and Integration in Healthcare for domain knowledge enrichment

[1]Subiksha.K.P, [2] M.Ramakrishnan
[1&2]School of Information Technology , Madurai Kamarajar University. India

ABSTRACT
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. 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 by 37% 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.
Keywords: Concept-based Information Retrieval, ATOMS (AGENT based Terminology Management System with Ontology Representation), ICR, WSD, CRT.

Received 22.03.2015 Revised 26.03.2015 Accepted 31.03.2016

INTRODUCTION

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. Early approaches by Voorhees [3] used general lexical thesauri such as WordNet for the purposes of query elaboration. WordNet is large general English language ontology where all the Nouns, verbs adjectives and adverbs are grouped into relative synonyms each expressing a distinguishable concept [4].

Concept search techniques were developed because of restrictions imposed by recognized Boolean keyword search technologies especially in dealing with large, unstructured digital collections of text [5]. 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. 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 [8]. 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.

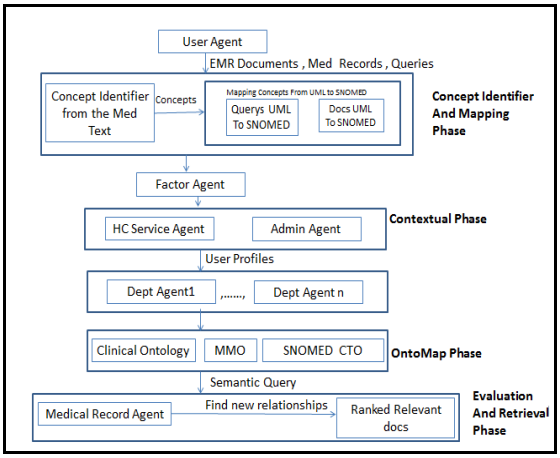

Figure 1 - ATOMS Architecture.

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.

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, modeled 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 authorized agents may access and update the EMR.

III. 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.

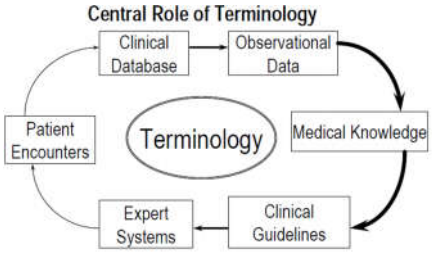

Figure 2 – Role Of Terminology

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.

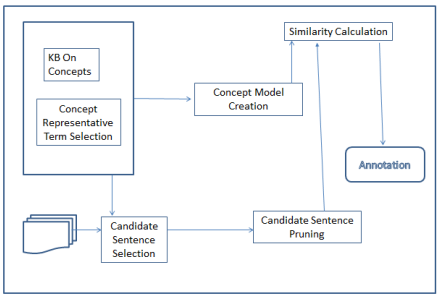

Figure 3 - Components of the Proposed Solution

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 [13] and matrix Jcn [14] which are proven to perform well among the unsupervised algorithms in paraphrase identification task [15]. 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 [16] 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 [17], as well as a predict vector-based measure Word2vec [18] 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.

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

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.

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

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).

RESULTS

MIMIC II database (Saeed et al., 2002) http://mimic.physionet.org is used for testing and development purpose. It consists of discharge summaries, electrocardiogram, echocardiogram, and radiology reports. 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 1 - 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 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 [26].
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. 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.

Table 2 - MAP and Precision Measures

Approaches	**Mean Avg Precision**	**Precision**
Keyword baseline Approach	0.2124	0.2743
Concept-based Approach	0.2352	0.3462
Role and Ontology based Approach	0.3717	0.4713

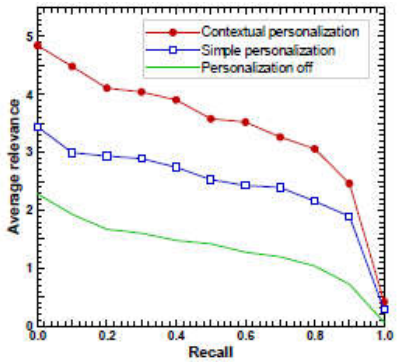

Figure 4 - Comparative performance of personalized search with and without contextualization averaged over ten use cases

For this purpose, it would be 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). This requires building a test bed consisting of a search space corpus, a set of queries, and a set of hypothetic context situations, where users would be required to provide ratings to measure the accuracy of search results. The latter means considering sequences of user actions defined a priori, which makes it more difficult to get a realistic user assessment, since in principle the user would need to consider a large set of artificial, complex and demanding assumptions.

Table 3 - Precision of context vector based WSD

Word Sense Disambiguation Approach	**Prec**
Random	51.73
Most Frequent	49.6
Context Vectors	87.25
Naïve Baiyes (NB)	93.46
TF-IDF CCV (Context Concept Vectors)	87.54

Table 4 - Precision of Ontology Graph based WSD

Word Sense Disambiguation Approach	**Prec**
OptimalDist	68.24
Ontology Shortest Path	77.78
Nearest Neighbour (NN)	74.32

Have observe that such 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 WSD, for example using interval of OptimalDist measures or Nearest-Neighbours(NN) on the ontology graph[23].

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

CITATION OF THIS ARTICLE

Subiksha.K.P, M.Ramakrishnan .A Comprehensive Approach for Knowledge Acquisition and Integration in Healthcare for domain knowledge enrichment. Bull. Env. Pharmacol. Life Sci., Vol 5 [5] April 2016: 01-07