

Domain Ontology Construction from Biomedical Text

Saurav Sahay¹, Baoli Li¹, Ernest V Garcia², Eugene Agichtein³, Ashwin Ram¹

¹College of Computing, Georgia Institute of Technology, Ph -404.894.7865

²Department of Radiology, Emory University, Ph – 404.353.0143

³Department of Mathematics and Computer Science, Emory University, Ph – 404.727.7962

{ssahay, baoli, ashwin}@cc.gatech.edu, eugene@mathcs.emory.edu,

Ernest.Garcia@emoryhealthcare.org

Keywords: Ontology Learning, Information Extraction, Natural Language Processing

Saurav Sahay – Presenting author

(submitted to International Conference on Artificial Intelligence (ICAI), 2007)

Abstract

NLM's Unified Medical Language System (UMLS) is a very large ontology of biomedical and health data. In order to be used effectively for knowledge processing, it needs to be customized to a specific domain. In this paper, we present techniques to automatically discover domain-specific concepts, discover relationships between these concepts, build a context map from these relationships, link these domain concepts with the best-matching concept identifiers in UMLS using our context map and UMLS concept trees, and finally assign categories to the discovered relationships. This specific domain ontology of terms and relationships using evidential information can serve as a basis for applications in analysis, reasoning and discovery of new relationships. We have automatically built an ontology for the Nuclear Cardiology domain as a testbed for further enhancing our techniques.

1. Introduction

One of the broad goals of intelligent information processing is domain modeling and knowledge creation for machine processable knowledge-aware applications. Biomedical informatics research has been actively looking at incorporating domain knowledge in ontologies that can be shared by many applications. Various ontologies and knowledge bases have been developed for several domains. UMLS[1] is one such large consolidated repository of medical terms and their relationships, spread across multiple languages and disciplines by combining more than 100 different source vocabularies. This ontology organizes information from various source vocabularies, each with its attributes, describes relationships between concepts and organizes these concepts under a semantic network of broad category types and links. However, UMLS ontology does not incorporate complex domain relationships between its resources. For example, although UMLS contains details about many diseases, viruses and bacteria, it does not incorporate relations between diseases and the causes of the diseases.[2]

We propose an approach of constructing a domain-specific ontology directly from a corpus using statistical NLP techniques. The discovered relationships between concepts are used to build a network of terms and relationships for the domain. We call this network our context map. We link the context map with the UMLS network by the concept matcher process to discover similar concepts based on lexical and conceptual metrics of similarity among two concepts. We create richer semantics in our ontology by explicitly capturing the type of the relationship between concepts. The advantage of this approach is that it directly uses the

domain data to construct smaller and manageable domain specific ontologies that can be used directly for several knowledge processing tasks.

Some important related work in this area point to TextToOnto (Maedche and Staab, 2000; Cimiano et al., 2005) and OntoLT (Buitelaar et al., 2004). Knowledge Engineering based approaches in ontology modeling assisted with Protégé or GATE are also quite popular.

2. Domain ontology construction framework

This section explains our technique for discovering domain-specific terms and relationships for constructing the ontology. We show the flow diagram and the techniques used therein in Table 1.

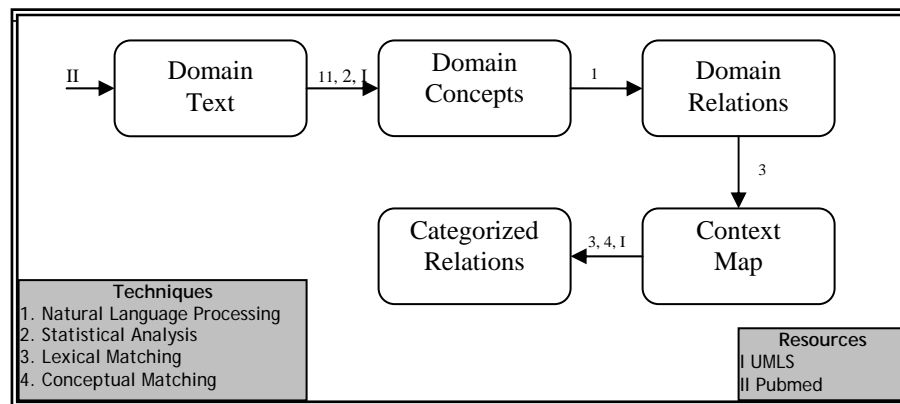


Table 1. Flow Diagram for our Ontology Construction System

2.1 Concept Extraction

Our goal is to build domain ontology for the Nuclear Cardiology discipline that deals with clinical trial experiments and does not involve animal subjects. We have built an information processing infrastructure which utilizes Pubmed web services to fetch desired abstracts from the Journal of Nuclear Cardiology and convert them to xml format. Our system retrieves abstracts from Pubmed and invokes the Mmtx system[3] to map all possible phrases in the abstracts to their possible UMLS concept identifiers and Semantic Types. Noun Phrases from titles, abstracts and associated Mesh Terms are extracted using MedPostSKRTagger[4], a Part of Speech tagger bundled with UMLS resources, trained on Pubmed corpus. Approximate matching heuristics such as occurrence of gaps and permutations of words in retrieved phrases are applied to look up the variants, synonyms and abbreviations for matching domain terminology with UMLS concepts. A higher weight is given to terms occurring in Titles compared to terms occurring elsewhere. A stopwords list used in SMART [5] system is used to discard non-informative words from our extracted phrases. Term frequencies are used as the keyword weighting scheme for ranking our phrases. A list of ranked retrieved phrases with their matching UMLS concept identifiers and Semantic Types is constructed. To build a clean ontology, we have to assign the best-matching concept identifier to the phrase which may map to several possible concept identifiers.

2.2 Context Extraction

The next step is to build a context map of related terms for our domain. This is done using the MedPostSKR Tagger, which is used to retrieve NounPhrase verbPhrase NounPhrase (Subject verb Object (SVO) triple patterns) patterns from sentences.

```
For each abstract A
  For each sentence S in A
    Find occurrences of domain concept pairs in S → Pairs
    For each concept pair <C1, C2> in Pairs
      Apply verbGroup matching classify <C1,C2> into relations → R
    Add R to all Relations
return Relations
```

Figure 1: Relation extraction algorithm

This above pseudo-code in Figure 1 is applied to find SVO patterns that map to one or many concept identifiers in UMLS. Categories of verbGroups are created using regular expression patterns for identifying a rich set of labeled relationships. GraphViz tool [6] is used to visualize the relationships between concepts.

2.3 Concept Matcher Process

2.3.1 Semantic Match

This step finds best matching concept from UMLS and assigns them to the domain phrase along with the match value. In order to find the conceptual match between the phrase and the possible concepts, the context map is used to find related phrases. These related phrases can map to a unique UMLS concept or many concepts. We look for related phrases with uniquely mapping concepts and then build a suffix tree for the concept hierarchy compute the conceptual distance [7] between the possible phrase concepts and the unique concept. The shortest distance between two concepts is computed by a spreading activation search on the UMLS hierarchy graph. Two concepts share a common concept in the graph. There may be several common concepts between two concepts. (through various parents of these two concepts) The shortest path common concept links the closest matching concepts. This notion of conceptual similarity is based on the premise that “*related terms in the context map are closely related to each other in the UMLS graph*”.

```
For each phrase P with multiple matching concepts <C1,C2 ... Cn>
  For each related concept C from the context map
    Build concept graph using PARENT links as a suffix tree
    Find TS(C,C1), TS(C,C2)... TS(C,Cn)
    shortestDist = min{ TS(C,C1), TS(C,C2)... TS(C,Cn)}
  return shortestDist
```

Figure 2: Concept Matcher algorithm

Figure 2 above describes the pseudo-code for finding concept identifier for terms with ambiguous concepts.

The tree similarity TS between two concepts C_1 and C_2 , can be computed according to the following formula [8]:

$$TS(C_1, C_2) = \frac{2 \times \text{Common}(C_1, C_2)}{\text{Depth}(C_1) + \text{Depth}(C_2)}$$

where $\text{common}(C_1, C_2)$ denotes the number of common nodes in the paths between the root and the given concepts, and $\text{depth}(C)$ is the number of nodes in the path connecting the root and the given concept C .

2.3.2 Lexical Match

We may not find a semantic match between arbitrary concepts in UMLS. ‘It should be pointed out that UMLS defined a parent Concept Unique Identifiers (CUI) only for a minority of CUIs, usually mutating the parents from the titles of classification sections (e.g. "Bronchial-Diseases").[9]’ Other links such as ‘broader’ or ‘narrower’ in the Metathesaurus are not well defined as they reference related terms from different vocabularies and can contain cycles and other ambiguities.

In order to be able to assign a unique class label to all our domain terms, we define another Lexical Match metric to assign the best possible categories to our terminology :

$$\text{LexicalMatch} = \text{Max} \left[0, \frac{\text{Num}(\text{Phr} \cap \text{Cpt}) - (\text{Num}(\text{Phr} - \text{Cpt}) + \text{Num}(\text{Cpt} - \text{Phr}))}{\text{Max}(\text{Phr}, \text{Cpt})} \right]$$

where we subtract the unique tokens from the common tokens and normalize by the maximum length token. This metric is suitable for multi word tokens (phrases and concepts) and computes the lexical similarity between words.

2.4 Relationship Matcher Process

A user might be interested in certain types of relationships. It is necessary to assign types to the retrieved verbs in relationship triples. We have created patterns for the 54 UMLS Semantic Network relationship links. We assign these categories to the extracted verbs. In order to increase the accuracy of this technique, we have built a system that uses WordNet resource for finding all word sense synonyms for the verb to match against our patterns list.

```

patterns = List of patterns for Semantic Network relationships
synonyms = List of synonyms for extracted domain verb
initialize a HashMap of patternMatches
  for each s in synonyms {
    patternMatches = longestStringMatch(s, patterns)
  }
return patternMatches

```

Figure 3: Relation Matcher algorithm

3. Experiments

3.1 Concept Extraction

The abstracts from 618 original scientific articles published in the Journal of Nuclear Cardiology (JNC) were used for this research and experiments. These abstracts summarize

the knowledge of the entire scientific manuscripts published by JNC from 1995 to 2004. We have used this resource to build the knowledgebase of relevant terms and relations. We extracted 10191 term (phrases) from these articles. These terms are noun phrases that have at least one mapping in the UMLS metathesaurus.

We used an expert provided list of 42 relevant domain terms for the Myocardial Perfusion Imaging domain (a sub-domain of nuclear cardiology). Of these, we found exactly matching 31 terms, 2 morphological variants and 8 phrases that contained the terms from the expert list. We missed to report 1 term in our ontology – ‘dyperidamole’ – which is a drug given during pregnancy. We found that this term didn’t exist in the 618 JNC abstracts we were using. Instead, another term, doperidamole (spelling variant) existed in our term list. To do away with noise in our data, we calculated the average frequency (2.9) of the 10191 retrieved terms and set that as our threshold. We pruned the list to 1415 terms – this list included all the 41 relevant domain terms or their variants.

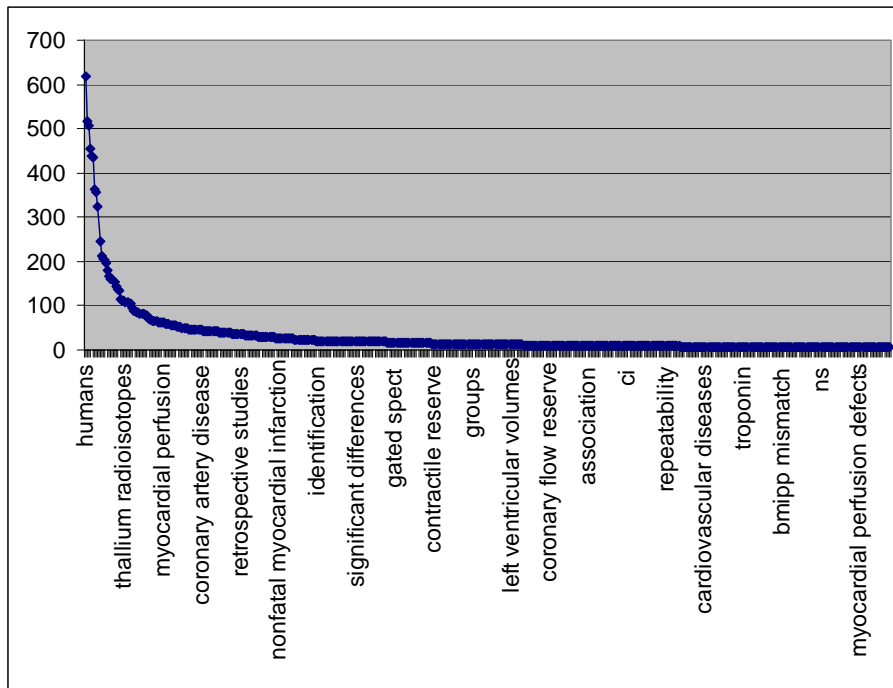


Figure 4: Top 500 Word Frequency Distribution from extracted terms

3.2 Context Extraction

Many complicated and expensive techniques have been applied extensively in literature to extract relationships and dependencies from data. These techniques generally rely on a deep parse results that gives a rich representation which is used for applications in knowledge extraction and semantics. The problem with biomedical data parsing is that there is no exhaustive tagged training corpus available to build models for efficient parsing on test data. Also, deep parsing is computationally very expensive compared to shallow parsing that uses the finite state automaton techniques. UMLS graciously provides such a shallow parser[4] that is trained on entire Medline corpus and claims to achieve over 97% accuracy.

We ran the shallow parser on our JNC abstracts and extracted NounPhrase verb NounPhrase triples for the 1415 extracted terms in our domain. We ignored non-informative verbs ('be',

'have' modal verbs) to extract meaningful relationships. We checked for all possible relations in all the sentences of the corpus. We retrieved 2562 such relations.

Some example relations extracted from our technique :

Left ventricle DIVIDED regional myocardial uptakes Acipimox REDUCES serum free fatty acid Mental stress FOUND significant hemodynamic responses Tl-201 washout rate CALCULATED lv end-diastolic volume index Tl-201/bmipp subtraction polar map SHOWED focal uptake pattern Carvedilol IMPROVES left ventricular function
Figure 5. Extracted Relations using a shallow parser

The goal of this stage was to build local contextual maps of terms and their relationships with other terms. We merged all the extracted relations and created visual maps for facilitating interpretation and analysis.

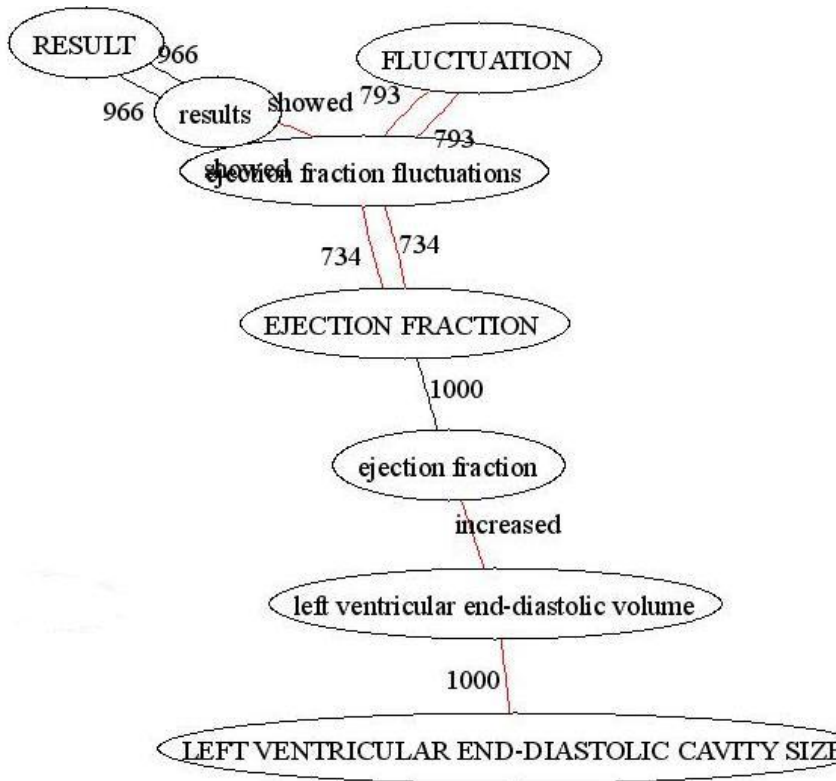


Figure 6. Part of context map with ranks and semantic types.

This context map shows two relationships {ejection fraction increased left-ventricular end-diastolic volume} and {results showed ejection fraction fluctuations}. The domain terms in

this map are attached to their possible semantic labels and a similarity score between the term and the concepts.

3.3 Concept Matcher Process

The goal of this step is to assign unique categories to the domain terms. One term in UMLS can map to several Metathesaurus categories depending on its lexical match (overmatch, undermatch, match with gaps), acronyms, variant forms, and its context information.

In order to disambiguate this, we have used a combination of semantic and lexical match to assign best possible categories to the domain terms in the ontology. Paper[4] has described conceptual match experiments. We ran into several problems with conceptual similarity metric as the concept tree grows exponentially in size following the concept node's parent links. We had to limit our searches to 500 parent nodes and search in this region of graph for the related term's concept hierarchy matches. We implemented a suffix tree based graph search algorithm to quickly scan for nodes in the suffix tree. We found very few matches with our limited implementation that we report here and we leave further discussions on this technique for future work.

We found "i-123 mibg" correctly mapping to "3-Iodobenzylguanidine" using the related node "analysis aspect". Similarly, 'nonfatal myocardial infarction' mapped to 'myocardial infarction' using 'unstable angina'

We assigned most categories to the domain terms using Lexical Match Metric.

```
intracoronary sestamibi <C1337333:SESTAMIBI>
endothelium-dependent regulation <C0851285:Regulation>
left ventricular end-diastolic volume <C0455833:Left ventricular end-
diastolic cavity size>
17-segment model <C0026336:Study models>
body <C0242821:Human body>
opposite the difference <C0443199:Differential quality>
successful evaluation <C1550157:Processing type - Evaluation>
thrombolytic therapy <C0040044:Thrombolytic Therapy>
diabetic nephropathy <C1442864:DIANPH gene>
cox regression analysis <C0034980:Regression Analysis>
```

Figure 7: Categorizing Domain Terms

3.4 Relationship Matcher

Figure 3 describes our Relationship Matcher algorithm. UMLS assigns 54 relationship types in its Semantic Network. We have used those label names had-crafted more patterns for those relationships. We check for verbs and their verbal synonym synsets and match against all patterns to assign the largest string match pattern as its category. This way, we have found several 'OTHER' patterns that do not belong to any of the 54 link patterns. On studying these, we have mostly found these 'OTHER' patterns to be noise in our data and safely discarded them. After all these steps, we have eventually found 1150 relationships between our domain concepts.

```

left ventricle DIVIDED <PART_OF> regional myocardial uptakes
study EVALUATED <MEASURES> impact
prognosis population COMPRISED <CONTAINS> 16,020 consecutive patients
cardioinhibitory response SHOWED <EXHIBITS> vasodepressor response
eighteen patients UNDERWENT <BRINGS ABOUT> i-123 mibg
constant supply SUSTAIN <PREVENTS> contractile function
reversible regional wall motion abnormalities PREDICT <CAUSES> future ca

```

Figure 8: Relation Matcher Results

4. Conclusion

This paper introduced an on-the-fly ontology construction methodology from text using existing resources and parsing and extraction. Using this elaborate framework and techniques, we have automatically and efficiently discovered relationships between resources.

We are working with experts in Nuclear Cardiology to extensively evaluate our results. We have been encouraged to learn about our promising results. As future work, we are looking at several techniques to refine and enrich our ontology. We are also in the process of converting our ontology to RDFS format. Once we have this machine processable knowledgebase in place, we can easily extend this to an environment where we can carry out knowledge reuse, decision support and reasoning for biomedical applications.

References

- [1] <http://umlsks.nlm.nih.gov/>
- [2] Sougata Mukherjea and Saurav Sahay. Discovering Biomedical Relations Utilizing the World-Wide Web. Pacific Symposium on Biocomputing 11:164-175(2006)
- [3] <http://mmtx.nlm.nih.gov/>
- [4] L. Smith, T. Rindfleisch, and W. J. Wilbur. Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14), 2004.
- [5] Buckley, C., Salton, G., and Allan, J. 1993. The SMART information retrieval project. In Proceedings of the Workshop on Human Language Technology (Princeton, New Jersey, March 21 - 24, 1993). Human Language Technology Conference. Association for Computational Linguistics, Morristown, NJ, 392-392.
- [6] Gansner, E. R. and North, S. C. 2000. An open graph visualization system and its applications to software engineering. *Softw. Pract. Exper.* 30, 11 (Sep. 2000), 1203-1233.
- [7] Caviedes, J. E. and Cimino, J. J. 2004. Towards the development of a conceptual distance metric for the UMLS. *J. of Biomedical Informatics* 37, 2 (Apr. 2004), 77-85.
- [8] I Spasic, S Ananiadou, J McNaught, A Kumar. Text mining and ontologies in biomedicine: Making sense out of Raw Text. Briefings in Bioinformatics, 2005 - Oxford Univ Press
- [9] DM Pisanelli, A Gangemi, G Steve 1998. An Ontological Analysis of the UMLS Metathesaurus. Proc AMIA Symp, 1998
- [10] Q. Li, P. Shilane, N. F. Noy, & M. A. Musen. Ontology Acquisition from On-line Knowledge Sources In the Proceedings of the AMIA Annual Symposium, Los Angeles, CA.
- [11] Maedche, A. and Staab, S. 2001. Ontology Learning for the Semantic Web. *IEEE Intelligent Systems* 16, 2 (Mar. 2001), 72-79.
- [12] Kietz, J., Volz, R., and Maedche, A. 2000. Extracting a domain-specific ontology from a corporate intranet. In Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning - Volume 7 (Lisbon, Portugal, September 13 - 14, 2000). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 167-175.