# Semantic Annotation and Inference for Medical Knowledge Discovery

Saurav Sahay
College of Computing
Georgia Institute of Technology
ssahay@cc.gatech.edu

Eugene Agichtein
Mathematics & Computer Science
Emory University
eugene@mathcs.emory.edu

Baoli Li
Math & Computer Science
Emory University
baoli@mathcs.emory.edu

Ernest V. Garcia
Department of Radiology
Emory University
Ernest.Garcia@emoryhealthcare.org

Ashwin Ram
College of Computing
Georgia Tech
ashwin@cc.gatech.edu

## Abstract

*We describe our vision for a new generation medical knowledge annotation and acquisition system called SENTIENT-MD ("Semantic Annotation and Inference for Medical Knowledge Discovery"). Key aspects of our vision include deep Natural Language Processing techniques to abstract the text into a more semantically meaningful representation guided by domain ontology. In particular, we introduce a notion of* semantic fitness *to model an optimal level of abstract representation for a text fragment given a domain ontology. We apply this notion to appropriately condense and merge nodes in semantically annotated syntactic parse trees. These transformed semantically annotated trees are more amenable to analysis and inference for abstract knowledge discovery, such as for automatically inferring general medical rules for enhancing an expert system for nuclear cardiology. This work is a part of a long term research effort on continuously mining medical literature for automatic clinical decision support.*

## 1. Introduction

The rapidly increasing volume of unstructured biomedical information poses the challenge of efficient and automated knowledge understanding so as to build autonomic computing systems that can acquire, represent, learn and maintain such knowledge, and efficiently reason from it to aid in knowledge discovery and re-use. The construction of these automated systems to assist biomedical decision making is impeded by difficulties in formalizing knowledge and in encoding that knowledge for use by computer agents that can integrate and reason from it.

Automatic semantic markup of unstructured textual data in the medical domain is also a challenging task. In particular, one of the challenges is the appropriate level of abstraction on which to annotate text. In the context of our vision of recognizing and reasoning about *general* knowledge in the findings of medical literature, we must appropriately annotate and disambiguate the concepts and relationships expressed in text as a first crucial step in this process.

We present our current work on a new semantic annotation and inference platform called SENTIENT-MD for precise semantic annotation of medical knowledge in natural language text. Our approach is to semantically annotate natural language parse trees, transforming them into annotated *semantic networks* for the purpose of inferring general knowledge from the text.

SENTIENT-MD is part of a larger project called *MERLIN* ('Medical Rule Learning') that aims to use domain corpus to generate domain-specific, evidence-based, machine processable knowledge that can be incorporated into knowledge based expert systems for the diagnosis and prognosis of coronary artery disease in patients imaged with ECG-gated myocardial perfusion (SPECT). We have developed an initial prototype system that uses SOAP web services calls to identify relevant abstracts from PubMed database and ranks and classifies the articles at the abstract and sentence level into different categories to aid in rule extraction from sentences that contain rule like knowledge in them.

## 2. Related Work

To the best of our knowledge, we have not seen similar work on semantic tree merging and rule extraction system. However, there have been several related work on semantic role labeling [14], [11], [6], textual inference [4], [7], dependency parsing [5] and ontology alignment [8]. For example, the textual inference task is to determine if

the meaning of one text can be inferred from the meaning of another and from background knowledge. Relationship Extraction system also apply heuristics, path learning and parsing techniques [15], [1]. The relationship extraction system aim at finding pre-determined paths and then apply a machine learning algorithm to learn such unseen paths. Our approach here uses simple yet high precision scoring functions for appropriate tree merging and creates a robust graph based infrastructure for semantic analysis and inference.

## 3. System Description

The overall pipeline for our prototype Rule Extraction system from Nuclear Cardiology abstracts using machine learning, language understanding and statistical techniques is described in Figure 1. The focus of this paper lies in taking the knowledge rich sentences and converting them to a form amenable to complex tasks like rule extraction and knowledge mining.

### 3.1. Dependency Parsing

The first stage in our process is to create dependency graphs for the sentence, which can be viewed as a complete tree structure containing triple information about every two connected token in the tree. As basis for the semantic graph, we use typed dependencies, (output by the Stanford parser [9]. in which each node is a word (the governor and the dependent) and labeled edges represent grammatical relations between words. We generate the dependency tree from the dependency relation vectors by joining the nodes of the relations. The root of this dependency tree is the head of the parse tree. In case the head of the parse tree is an auxiliary verb, we make it's governing dependency relation as the head of the dependency tree. Figure 2 gives the typed dependency graph for a simple sentence 'SPECT uses radioactive tracers for imaging' before and after our merging procedure. The semantic graph for a sentence contains thus a node for each word of the sentence. These nodes contain the governor-dependent relation along with the type of relation labeled as the directed arc between the two.

We have used the JUNG (Java Universal Network/Graph)(`http://jung.sourceforge.net/`) which is an open-source software library that provides a common and extendible language for the modeling, analysis, and visualization of data that can be represented as a graph or network. It provides algorithms for finding the paths between nodes of the tree along with several other graph theoretic algorithms.

### 3.2. Annotation

Like the Resource Description Framework (RDF) language for the Semantic Web which provides a simple data model for describing relationships between resources in terms of named properties and their values, our system can represent biomedical resources such as diseases, drugs, procedures, equipment, etc. as well as general relationship type concepts such as 'cause', 'treat', 'part of', etc as described in the UMLS Semantic Network [10]. Based on pattern matches, we have identified simple instances of date and time as resources for annotation as well. Our system invokes the Mmtx system(`http://mmtx.nlm.nih.gov/`) to map all possible phrases in the abstracts to their possible UMLS concept identifiers and Semantic Types. Noun Phrases and Preposition phrases from abstracts are extracted using MedPostSKRTagger[13], a Part of Speech tagger bundled with UMLS resources, trained on Pubmed corpus. A stopword list used in SMART [2] system is used to discard non-informative words from our extracted phrases for the annotation system.

### 3.3. Merging

The novelty of our system comes from application of this Merging step to our annotated dependency trees. We apply state of the art distance metrics to compute similarity between ontological concepts, phrases and nodes in the tree. The Jaro distance metric[3] takes into account typical spelling deviations and has been shown to perform well in probabilistic record linkage literature. For two strings s and t, let s' be the characters in s that are 'common with' t, and let t' be the characters in t that are 'common with' s; roughly speaking, a character a in s is 'in common' with t if the same character a appears in about the place in t.

Jaro-Winkler metric is an extension of the Jaro distance metric, which modifies the weights of poorly matching pairs s, t that share a common prefix. This adjustment gives more favorable ratings to strings that match from the beginning for a set prefix length. The paper [3] describes a combination of token based distance function (TFIDF) and edit distance like similarity function (Jaro-Winkler) called Soft-TFIDF which has given highest overall performance in their studies over a range of different datasets.

In SoftTFIDF, similar tokens are considered as well as tokens in $S \cap T$. Secondary similarity function (Jaro-Winkler in our case) is combined with the standard TFIDF function to incorporate approximate matching in an efficient manner. A $CLOSE(\theta, S, T)$ function is defined as the set of words $w \epsilon S$ and some $v \epsilon T$ such that secondary distance $dist'(w, v) > \theta$, and for $w \epsilon CLOSE(\theta, S, T)$, $D(w, T) = max_{v \epsilon T} dist(w, v)$.

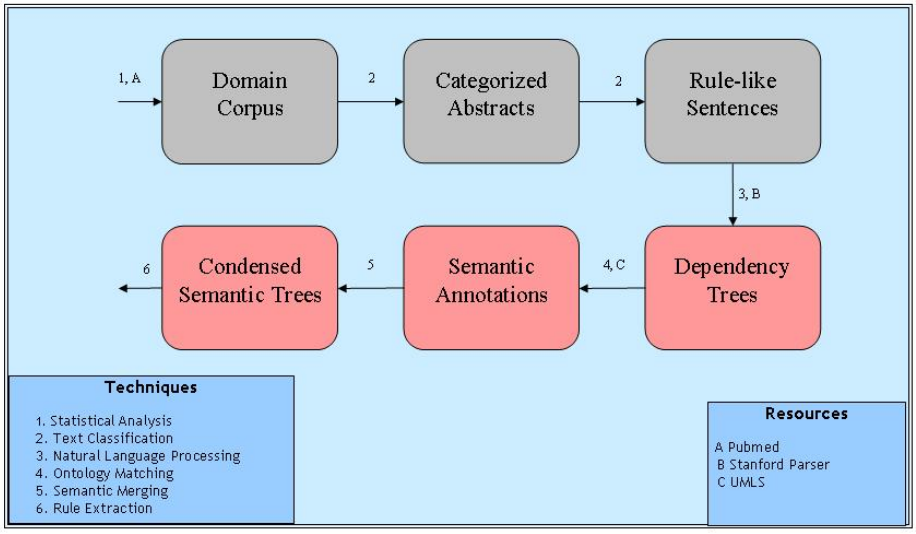


**Figure 1. Architecture of our system**

```
Input: Sentence, Ontology
Output: Merged Semantic Tree
<C> = matchConcept(sentence)
<D> = getDependencies(sentence)
Tree = createTree(head, <D>)
While(Tree does not change)
   Forall c in <C>
      if(Match(c,Tree)>= threshold t)
         Merge(c,Tree)
return Tree
```

**Figure 2. Semantic Merging pseudo-code**

$$SoftTFIDF(S,T) = \sum_{w \epsilon CLOSE(\theta,S,T)} V(w,S).V(w,T).D(w,T)$$

V(w,S) is the standard TFIDF formula in the SoftTFIDF equation.

Each phrase with a mapping UMLS concept may map to several concepts ($mappings_i$) from UMLS. We disambiguate the mapping of such phrases (disambiguated concept c) to their corresponding UMLS concepts using the SoftTFIDF metric as described above.

$$c = max_i\{SoftTFIDF(phrase, mappings_i)\}$$

Merging of dependency trees occur according to Figure 2:

$$Match(c, Tree) =$$

$$max\{SoftTFIDF((c, c \cap headTree), (c, head), (c, headTree))\}$$

In this formula, headTree is the subtree of the head node matching the concept and it's immediate children.

The tree is searched for the head node (head) containing the matching term. In our iterative Tree merging process, we compare $level_i$ and $level_{i+1}$ for node merging. Concept merging can occur as a result of merging the head, or the head and it's children, or the head and its children nodes containing the concept. The tree is then rearranged after merging.

## 4. Experiments

We have done some initial experiments using our approach for machine understanding of biomedical textual data. In particular, we have analyzed some condensed semantic trees to see it's usability for discovering new relationships, mapping concepts to their appropriate ontological representation and in future, directly extracting expert system rules and assertions.

Figure 3 demonstrates our approach using an example sentence *'SPECT uses radioactive tracers for imaging.'*We have used the Stanford parser to parse the knowledge-bearing sentences and produce typed dependencies from the sentences. We build a graphical framework of these dependency relationships so that we can do inference over these graphs for information processing. We annotate the concepts in the graph using the SoftTFIDF match criteria and the relationships in the graph as described in [12]. Figure 3 shows the initial parse tree, the dependency graph, the compressed semantic graph and the concept node annotations for the sentence. Our example sentence in 3 after semantic processing is a simplified rule assertion - there is a single path between the two identified concepts 'SPECT' and 'Di-

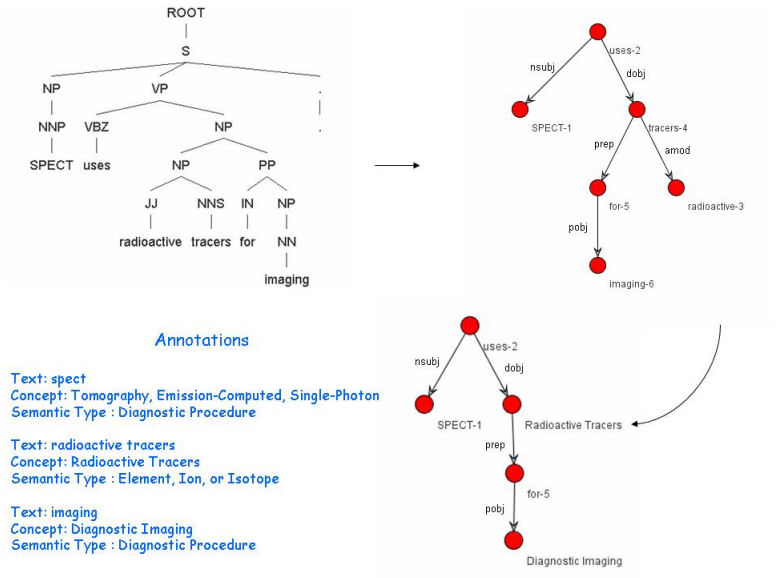


**Figure 3. From parse tree to compressed semantic graph**

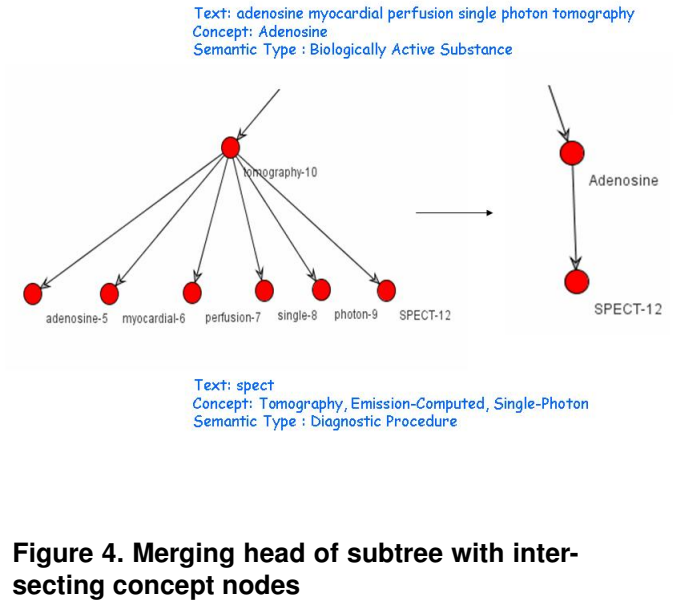


**Figure 4. Merging head of subtree with intersecting concept nodes**

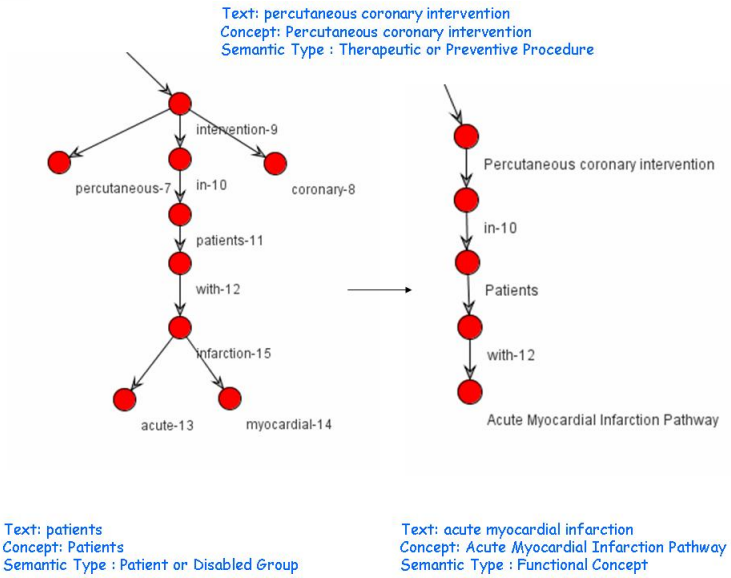


**Figure 5. Graph displaying our three merging criteria**

agnostic Imaging.'

Figure 4 displays a part of the semantic graph for the sentence *Preliminary studies indicate that adenosine myocardial perfusion single photon tomography (SPECT) can safely and accurately stratify patients into low and high risk groups early after acute myocardial infarction (AMI).*Experts have manually annotated this sentence as carrying an expert system rule in it. The graph demonstrates an example of our head merging criteria.

In particular, a maximal match is found between the noun phrase 'adenosine myocardial perfusion single photon tomography' and subtree containing those token nodes ($SoftTFIDF(C, C \cap headTree)$) resulting in the deletion of child nodes, nevertheless, retaining the same meaning of the sentence representation. Again using the SoftTFIDF match criteria, this phrase gets matched to the concept 'Adenosine' in the tree. The additional node 'SPECT' which has a relationship type 'abbreviation' with it's parent node 'tomography' retains the complete meaning of the phrase in tree.

Figure 5 demonstrates all of our merging criteria to compress a part of the semantic tree to a much simplified representation. The rule-bearing sentence used in this example is *Analysis of TF SPECT immediately after percutaneous coronary intervention in patients with acute myocardial infarction is a useful noninvasive method for evaluating coronary microvascular dysfunction.*Here, both singular conceptual entities 'percutaneous coronary intervention' and 'acute myocardial infarction' are merged into respective semantic entities, thus retaining the single logical path between entities to aid in extraction and inference.

## 5. Conclusion

We presented our vision for semantic annotation and inference to support discovery of general medical knowledge from published medical literature. Preliminary experiments and case studies in the nuclear cardiology domain indicate the promise of our approach. Our system is a work in progress, and we are actively experimenting with implementation alternatives. As continuation of this work, the semantic network representing the semantic and role relationships between the concepts will be constructed from our current sentence level semantic trees. Additional context-specific relationships between nodes will be inferred based on co-reference resolution and ontology relationships.

In particular, we observed that using the appropriate granularity of semantic annotation has significant implications for knowledge extraction. We have also found that user interfaces and visualization are crucial for obtaining reliable feedback from domain experts. We believe that the general approach and some of the techniques we described can serve as a robust platform for our long-term research into discovering and structuring medical knowledge from the literature.

## References


[1] E. Agichtein and L. Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, 2000.

[2] C. Buckley, G. Salton, and J. Allan. The smart information retrieval project. In *HLT '93: Proceedings of the workshop on Human Language Technology*, pages 392–392, Morristown, NJ, USA, 1993. Association for Computational Linguistics.

[3] W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string metrics for matching names and records.

[4] M.-C. de Marneffe, T. Grenager, B. MacCartney, D. Cer, D. Ramage, C. Kiddon, and C. D. Manning. Aligning semantic graphs for textual inference and machine reading. In *Proc. of the AAAI Spring Symposium at Stanford. 2007*, 2007.

[5] M.-C. de Marneffe, B. MacCartney, and C. D. Manning. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the IEEE / ACL 2006 Workshop on Spoken Language Technology*. The Stanford Natural Language Processing Group, 2006.

[6] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

[7] A. Haghighi, A. Y. Ng, and C. D. Manning. Robust textual inference via graph matching. In *HLT/EMNLP*. The Association for Computational Linguistics, 2005.

[8] Y. Kalfoglou and M. Schorlemmer. Ontology mapping: The state of the art. In Y. Kalfoglou, M. Schorlemmer, A. Sheth, S. Staab, and M. Uschold, editors, *Semantic Interoperability and Integration*, number 04391 in Dagstuhl Seminar Proceedings, 2005.

[9] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[10] A. McCray. Mccray a. an upper level ontology for the biomedical domain. comp functional genomics 2003; 4: 80-84., 2003.

[11] S. Pradhan, W. Ward, K. Hacioglu, J. Martin, and D. Jurafsky. Semantic role labeling using different syntactic views. In *ACL*, 2005.

[12] S. Sahay, B. Li, E. V. Garcia, E. Agichtein, and A. Ram. Domain ontology construction from biomedical text. international conference on artificial intelligence, icai 2006.

[13] L. Smith, T. Rindflesch, and W. J. Wilbur. Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321, 2004.

[14] K. Toutanova, A. Haghighi, and C. D. Manning. Joint learning improves semantic role labeling. In *ACL*, 2005.

[15] D. Zelenko, C. Aone, and A. Richardella. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106, 2003.