Semantator: A Semi-automatic Semantic Annotation Tool for Clinical Narratives

Dezhao Song[⋆,a,b], Christopher G. Chute[a], and Cui Tao[a]

[a]Division of Biomedical Statistics and Informatics, Mayo Clinic
200 First Street SW, Rochester, MN 55905
[b] Department of Computer Science and Engineering, Lehigh University
19 Memorial Drive West, Bethlehem, PA 18015
des308@cse.lehigh.edu,{chute,tao.cui}@mayo.edu

Abstract. In this paper, we introduce Semantator, a semi-automatic tool for document annotation with Semantic Web ontologies. With a loaded free text document and an ontology, users can annotate document fragments with classes in the ontology to create instances and relate created instances with ontology properties. Also, Semantator enables automatic annotation by connecting to the NCBO annotator and the clinical Text Analysis and Knowledge Extraction Systems (cTAKES). By representing annotations in OWL, Semantator has basic reasoning capability based upon the underlying semantics of *owl:disjointWith* and *owl:equivalentClass*.

Keywords: Semantic Annotation, Clinical Narratives, Semi-automatic

1 Introduction

Computerized approaches have been widely adopted to conduct clinical research, particularly by using data organized in a machine processable and understandable way. Manually converting free text based clinical documents to structured data is time-consuming. Automatic approaches that are based upon or adopting Natural Language Processing (NLP) techniques have been well studied [2, 1], but their performance may not always be satisfying. Structured data formats (e.g., Comma-separated Values (CSV), relational databases, eXtensible Markup Language (XML), etc.) have been well adopted; however, they do not enable data publishers to precisely embed the underlying semantics of their data.

In this paper, we introduce Semantator[1], a Protégé[2] plugin and a semi-automatic annotation tool for annotating clinical narratives using semantic web ontologies. Although Semantator is designed for annotating clinical documents, it can also be applied to documents in other domains. Figure 1 shows the main interface of Semantator. Currently, Semantator provides: 1) basic man-

⋆ This work was done when the first author was an intern at Mayo Clinic.
[1] http://informatics.mayo.edu/CNTRO/index.php/Semantator
[2] http://protege.stanford.edu

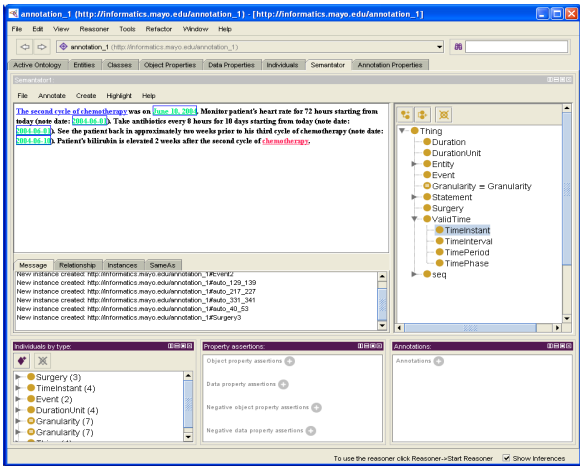

Fig. 1. Main Interface

ual annotation functionalities: ontology instance creation/deletion, relationship creation/deletion, linking equivalent instances and exporting/reloading existing annotations; 2) automatic annotation by connecting to the NCBO annotator [3] and cTAKES [4]; 3) basic reasoning support based on the underlying semantics of the *owl:disjointWith* and *owl:equivalentClass* predicates.

2 Manual Annotation

In Semantator, there are two ways to create instances: One-at-a-Time and Batch Creation. For the first option, a user can create one instance each time by selecting a piece of text in the loaded document and selecting a class from the loaded ontology in Protégé. The system allows the user to pick a color to be used for highlighting all instances of the selected class when this class is first used for annotation. The instance is associated with an rdfs:label: *<inst1,rdfs:label,[selected text]>* by default. A document may contain many instances that could actually be annotated with the same class and creating such instances individually can be time consuming, especially when the document is long. So, in Semantator, a user can add different document fragments into a candidate list, choose an ontology class and annotate all selected fragments to be instances of this chosen class. Annotated instances can be easily removed. The same document fragment could have been annotated to be instances of different classes. When a user chooses to delete the instance(s) of a document fragment, Semantator detects all associated instances. Then the user can choose to delete any of these instances individually.

In a clinical document, it is possible that instances that occur at different places in the document actually refer to the same real world entity[3]. For instance, in the following document, two instances (in bold) of the *Event* class have been created, and they actually represent the same event in real world.

[3] This is generally referred to as entity coreference [5].

The second cycle of chemotherapy was on June 10, 2004. Patient's bilirubin is elevated 2 weeks after ***the second cycle of chemotherapy.***

Such annotations on equivalences can be important to infer new knowledge and useful for medical care related applications [6]. Similar to instance batch creation, users can select an arbitrary number of annotated instances and add them to a *sameAs candidate list* to make them pairwisely equivalent.

Another important type of annotation is instance relationship annotation. Semantator allows users to create a single relationship between two instances at a time. They can select two instances and add them to the *relationship candidate list*. Then, they can choose any object property from the loaded ontology and decide the subject of this new relationship. A relationship can be easily deleted following a similar procedure as deleting an instance.

In Semantator, users can export their annotations to an RDF file together with a XML file that contains annotation related metadata, such as the position of each annotated instance, the color used to highlight instances of each class, etc. Next time when a user opens the same document, Semantator allows users to reload their previous annotations by choosing the correct RDF and XML files.

3 Semi-automatic Annotation

To facilitate the annotation process, we introduce the semi-automatic annotation capability of Semantator by adopting the NCBO annotator [3] and cTAKES [4].

BioPortal [3] is a Semantic Web based platform designed for the Biomedical domain, enabling users to search for specific ontologies that match certain user provided keywords. It provides an online annotation tool, the NCBO annotator, that recognizes relevant biomedical ontology terms with user-chosen ontologies in free text. Before starting annotation, we provide a list of ontologies (by querying BioPortal) that are currently supported by BioPortal and a user can choose an arbitrary number of ontologies from this list against which the annotator will match the words and phrases in a loaded document. All automatically annotated entities are highlighted in Semantator and users can choose to only retain those correctly identified instances from their perspective. cTAKES is another tool used in Semantator for automatic annotation in a similar way to that supported by NCBO annotator. Different from the NCBO annotator, cTAKES 1) is designed for clinical domain; 2) adopts NLP techniques and supports negation and time constraints. Currently, cTAKES performs annotation with the SNOMED CT and RxNorm dictionaries but users can add their own dictionaries as needed.

One advantage of annotating documents with ontologies is that users can benefit from the reasoning capabilities provided by Semantic Web techniques. Semantator currently supports two types of reasoning based upon class disjointness and class equivalence. Using the automatic annotation services, the same document fragment might be annotated to be candidate instances of disjoint classes. Take the following sentence as an example:

I was pleased to inform Mr. Smith that his PSA ***today*** *is undetectable.*

By calling the NCBO annotator with the SNOMED CT ontology, the word *today* is annotated to be an *Organic Chemical*; but a human may simply annotate it to be an *TimeInstant* from the CNTRO ontology [7]. Assuming we have the knowledge about the disjointness between the two classes: *Organic Chemical* and *TimeInstant*, Semantator will report an inconsistency. Similarly, if we annotate this sentence with both NCI Thesaurus and the International Classification Nursing Practice (ICNP) ontologies, BioPortal will annotate the word *today* to be an instance of the *Antibiotic* class in both ontologies. If we assume that the two *Antibiotic* classes from the two ontologies are equivalent and a user only annotated *today* to be an instance of one of them, Semantator will then suggest the user to also annotate it to be an instance of the other.

For future work, we are developing a DIFF module to visualize the differences and calculate the inter-annotator agreement between annotations of different annotators. Also, we would like to enhance Semantator with some query capability for users to search within the annotation results. Finally, we will explore how to provide a more general reasoning capable framework within Semantator.

Acknowledgement

This research is partially supported by the National Center for Biomedical Ontologies (NCBO) under the NIH Grant #N01-HG04028, and the NSF under Grant #0937060 to the CRA for the CIFellows Project.

References

1. Aronson, A.R., Lang, F.M.: An overview of metamap: historical perspective and recent advances. JAMIA 17(3), 229–236 (2010)
2. Demner-Fushman, D., Chapman, W.W., McDonald, C.J.: What can natural language processing do for clinical decision support? Journal of Biomedical Informatics 42(5), 760–772 (2009)
3. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A.D., Chute, C.G., Musen, M.A.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Research 37(Web-Server-Issue), 170–173 (2009)
4. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 17(5), 507–513 (2010)
5. Song, D., Heflin, J.: Domain-independent entity coreference in RDF graphs. In: Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM). pp. 1821–1824 (2010)
6. Tao, C., Solbrig, H.R., Sharma, D.K., Wei, W.Q., Savova, G.K., Chute, C.G.: Time-oriented question answering from clinical narratives using semantic-web techniques. In: 9th International Semantic Web Conference (ISWC). pp. 241–256 (2010)
7. Tao, C., Wei, W.Q., Solbrig, H.R., Savova, G., Chute, C.G.: CNTRO: A semantic web ontology for temporal relation inferencing in clinical narratives. In: American Medical Informatics Assocication Annual Symposium (AMIA). pp. 787–91 (2010)