## Analyzing Tools for Biomedical Text Annotation with Multiple Ontologies

Kele T. Belloze<sup>1,\*</sup>, Daniel Igor S. B Monteiro<sup>2</sup>, Tulio F. Lima<sup>2</sup>, Floriano P. Silva-Jr<sup>1</sup>

and Maria Claudia Cavalcanti<sup>2</sup>

<sup>1</sup> Laboratório de Bioquímica de Proteínas e Peptídeos, Instituto Oswaldo Cruz - FIOCRUZ, Rio de Janeiro, Brasil <sup>2</sup> Departamento de Ciência da Computação, Instituto Militar de Engenharia, Rio de Janeiro, Brasil

Textual databases such as PubMed, which contain references to scientific articles, are a major source for the extraction of useful information, since many scientific discoveries are deposited only in text form. However, due to the massive size of these bases, computational approaches are needed to extract information from texts. Currently, PubMed has more than 20 million citations from biomedical literature.

Ontology-based semantic annotation is an approach that aims to enrich the text with semantic descriptions and thereby facilitating the extraction of information based on semantic content embedded there. The ontologies used are usually focused on one domain. However, scientific articles often deal with different areas. For example, texts whose theme is about drug targets have references on molecular biology, pharmacology, chemical compounds and organism names. This suggests that the semantic annotation has to be done with multiple ontologies, to cover all or most of the domains of the text. Our motivation to annotate semantically with multiple ontologies is to be able to identify extra annotations that would be made manually. For example, let us imagine that in a text it is annotated the name of a gene with the Gene Ontology, the organism with NCBITaxon and the pharmacogenomic relationship of interest (such as the knockout technique) with the PHARE ontology. Realizing that the text mentions the knockout technique applied to the gene G of the organism O, causes its death, it would be useful then to annotate manually that G is essential for O.

To support semantic annotations, there are many tools available all over the web. This paper aims to identify and compare these tools, focusing on texts and ontologies in the biomedical area using two of main characteristics: 1) form of annotation and 2) flexibility in the ontology load. The form of annotation can be automatic or manual. In this study, we investigated automated tools to check if they have the option of manual annotation. The selection of only automatic tools is due to the large volume of texts and also because of the difficulty and high cost to keep specialists responsible for the task of manual annotation. The manual annotation additional feature is important because it would be used to insert extra annotation. With respect to the load flexibility of ontologies, some items are observed such as the size and format of the ontologies and the possibility of using an arbitrary ontology (a user choice). The utilization of arbitrary ontologies allows different domains to be used for the annotation. The tools with these characteristics have also been tested. Other characteristics observed involve: representation of the annotation (intrusive – in the text or non-intrusive – as attached file), types of documents compatible (which file extensions are supported as an input in the tool), documentation availability and platform of development (web or desktop).

Despite the fact that tools such as KIM, Ontea and RDFace generate automatic annotations, they have their own ontologies which are not on the biomedical domain. Knowtator is a plugin for Protégé Server and only a few tasks of the annotation process are automated. MnM, GoNTogle and RDFa Editor tools perform automatic annotation and have flexibility in loading arbitrary ontologies, but could not be used due to support problems. The NCBO Annotator is a web service that annotates full texts using ontologies from biomedical domain available at NCBO BioPortal. However, it is not available for immediate usage, and demands the development of a client to that web service.

AutôMeta and GATE can perform automatic annotation of documents and also have flexibility in loading arbitrary ontologies. These are selected as tools to be tested and used for the purpose of our experiment on semantic annotation.

AutôMeta uses RDFa (an annotation language recommended by W3C), has a reasoner to infer new annotations and it supports the load of large ontologies such as Gene Ontology and NCI Thesaurus, among others. The texts for annotation must be in 'txt' format and annotation is made using an intrusive method. GATE is a tool for natural language processing. It is very solid and mature in the task of semantic annotation using the resources of language and processing. Its differential is on being able to load different extensions of documents (txt, pdf, doc, etc.). Additionally, it performs non-intrusive annotation, and archives them in 'xml' files. Ontologies are loaded as processing resources, which can happen very slowly in the case of large ontologies. Both tools have good documentations and are free.

Moreover, it is noteworthy that these tools works with a set of input texts, but only uses one ontology at a time. Therefore, it is possible to have texts annotated with multiple ontologies, but in separate files, generating a new volume of texts and many output files. The simultaneous annotation with multiple ontologies is still an unsolved problem.