
Domain Specific Information Extraction for Semantic Annotation

----

By

Zeeshan Ahmed
Charles University,Prague
University of Nancy 2,France
ahmed_shani2002@yahoo.com

Semantic annotation is a helpful technique to understand the under laying semantics of the document. It provides additional information in the form of metadata which then makes documents to be processed in an intelligent way. The problem with semantic annotation is that these annotations are not universal e.g. semantic annotation for a document in particular domain might have different meaning in other domain. Therefore, domain specific knowledge is used for semantic annotation and this domain specific information is provided by ontologies.

The main problem with Semantic Annotation is availability of ontology for the domain. Ontology comprises of concept and relationships. In an ontology, a concept may be atomic or defined by a set of properties. This set of properties classifies the concept with other concept in ontology. In this thesis, we present an approach that deals with semantic annotation using properties of concept in an ontology rather than simple instance matching technique currently available. In this approach, the document is analyzed for the purpose of identifying these properties using ontology. If the properties found in document match with properties of any concept in ontology, the document is annotated with that concept. In this way, documents are indexed according to these properties.

The main target of this thesis is to present approaches of how these properties can be extracted from documents; both for the purpose of semantic annotation and ontology building. To achieve this target, we present two different approaches to information extraction for Semantic Annotation; "Rule Based Approach" and "Dependency Based Approach. We present the comparative analysis of effectiveness of these two approaches on a small corpus.

This kind of semantic annotation is useful for the efficient answering of search queries, clustering, text summarization etc. We apply this semantic annotation approach on the corpus of recipe documents. In our domain, these annotations are used for recipe adaptation purpose. In adaptation, the purpose is to intelligently replace some ingredient with other ones to make an adapted recipe. Apart from our main target of Information Extraction, we also propose an Ontology for our domain of recipe document as well as a process of semantic annotation.