Towards Controlled Natural Language for Semantic Annotation

Brian Davis, Digital Enterprise Research Institute, National University of Ireland, Ireland
Pradeep Dantuluri, Digital Enterprise Research Institute, National University of Ireland, Ireland
Siegfried Handschuh, Digital Enterprise Research Institute, National University of Ireland, Ireland
Hamish Cunningham, University of Sheffield, UK

ABSTRACT

Richly interlinked metadata constitute the foundation of the Semantic Web. Manual semantic annotation is a labor intensive task requiring training in formal ontological descriptions for the otherwise non-expert user. Although automatic annotation tools attempt to ease this knowledge acquisition barrier, their development often requires access to specialists in Natural Language Processing (NLP). This challenges researchers to develop user-friendly annotation environments. Controlled Natural Languages (CNLs) offer an incentive to the novice user to annotate, while simultaneously authoring his/her respective documents in a user-friendly manner. CNLs have been successfully applied to ontology authoring, but little research has focused on their application to semantic annotation. This paper describes two novel approaches to semantic annotation, which permit non-expert users to simultaneously author and annotate meeting minutes using CNL. Finally, this work provides empirical evidence that for certain scenarios applying CNLs for semantic annotation can be more user friendly than a standard manual semantic annotation tool.

Keywords: Controlled Natural Languages, Metadata, Natural Language Interfaces, Ontology-based Information Extraction, Semantic Annotation

INTRODUCTION

The Semantic Web endeavors to bring machine-processable meaning to the content of webpages. It envisions the Web as a universal medium for data, information and knowledge exchange, creating an environment where intelligent software agents can travel freely between web resources, carrying out sophisticated tasks for users\(^1\). In order for the Semantic Web to become a reality, we need, as a \textit{prima inter pares}, semantic data. The process of providing semantic data is very often referred to as semantic annotation, because it frequently involves the embellishment of existing data, i.e. the text, with semantic metadata, which can subsequently describe the
associated text. Hence semantic annotation is one of the core challenges for building the Semantic Web.

Manual semantic annotation however is a complex and labored task both time-consuming and expensive often requiring specialist annotators or the subsequent training of such annotators. This may require (an arguably unnecessary) exposure to formal ontological description. Such formal data representation can act as a significant deterrent for non-expert users or organizations seeking to annotate resources as part of their daily activity, thus allowing them to fully benefit from the adoption of Semantic Web technologies. While (Semi)-automatic annotation tools attempt to remove this constriction, which is commonly known as the knowledge acquisition bottleneck, their application often requires access to specialists who can combine Natural Language Processing(NLP)/Machine Learning(ML) and Semantic Web ontology languages. Such specialists are costly and rare and furthermore the creation or acquisition of quality language resources to bootstrap such approaches may require significant investment, which for a small to medium enterprises may not be justifiable. Consequently, this challenges researchers to develop user-friendly manual annotation environments to support the knowledge acquisition process.

Controlled Natural Languages (CNLs) offer an incentive to the novice user to annotate, while simultaneously authoring, her respective documents in a user-friendly manner, yet at the same time shielding her from the underlying complex knowledge representation formalisms of ontology languages. “Controlled Natural Languages are subsets of natural language whose grammars and dictionaries have been restricted in order to reduce or eliminate both ambiguity and complexity.” The use of CNLs for ontology authoring and population is by no means a new concept and it has already evolved into quite an active research area (Smart, 2008). Furthermore, a natural overlap exists between tools used for both ontology creation and semantic annotation, for instance the Controlled Language for Information Extraction(CLIE) technology permits ontology creation and population by mapping both concept definitions and instances of concepts to an ontological representation using a CNL called CLoNE - Controlled Language for Ontology Editing (Funk et al., 2008). Despite such efforts, very little research has focused on applying CNLs to semantic annotation. The reader should note that there is a subtle difference between the process of ontology creation and population and that of semantic annotation. We describe semantic annotation as “a process as well as the outcome of the process”. Hence it describes i) “the process of adding semantic data or metadata to content given an agreed ontology and ii) it describes the semantic data or metadata itself as a result of this process”(Handschuh, 2005). Of particular importance here is the notion of the \textit{addition or association} of semantic metadata to content. 

\textbf{Latent Annotation}

As with any annotation environment, a major drawback is that in order to create metadata about a document, the author must \textit{first} create the content and \textit{second} annotate the content, in an additional a posteriori, annotation step. In the context of our application of CNL to semantic annotation, we seek to merge both authoring and annotation steps into one. This process differs from classic \textit{a-posteriori} annotation resulting in a new type of annotation which we call \textit{latent} annotation. Latent comes from the Latin word with identical spelling who’s etymology is derived from the Latin verb latere (lie hidden), a nod in respect to \textit{a-posteriori}(later, what comes after). 

\textbf{Towards a Definition of Controlled Language for Semantic Annotation}

We hereby refer to the application of CNL to semantic annotation as controlled (natural language) annotation which reflects how traditional CNL intersects but also differs from our approach, whereby: controlled (Natural Language) Annotation or Controlled Annotation is the application of CNL technologies to the process of semantic annotation. Controlled
Managing QoS Degradation of Component Web Services in a Dynamic Environment
[www.igi-global.com/article/managing-qos-degradation-of-component-web-services-in-a-dynamic-environment/203697?camid=4v1a](www.igi-global.com/article/managing-qos-degradation-of-component-web-services-in-a-dynamic-environment/203697?camid=4v1a)

Mobile Service Oriented Architecture (MSOA) for Businesses in the Web 2.0 Era
[www.igi-global.com/chapter/mobile-service-oriented-architecture-msoa/39170?camid=4v1a](www.igi-global.com/chapter/mobile-service-oriented-architecture-msoa/39170?camid=4v1a)